IRE Transactions

ON AUTOMATIC CONTROL

Volume AC-6 MAY, 1961 Number 2

IRE Papers Presented at the
JOINT AUTOMATIC CONTROL CONFERENCE
University of Colorado, Boulder, Colo.
June 28–30, 1961

TABLE OF CONTENTS

PUBLISHED BY THE

PROFESSIONAL GROUP ON AUTOMATIC CONTROL

IRE PROFESSIONAL GROUP ON AUTOMATIC CONTROL

The Professional Group on Automatic Control is an organization, within the framework of the IRE, of members with principal professional interest in Automatic Control. All members of the IRE are eligible for membership in the Group and will receive all Group publications upon payment of the prescribed fee.

Annual Fee: $3.00

IRE TRANSACTIONS®
on Automatic Control

George S. Axelby, *Editor*, Air Arm Division,
Westinghouse Electric Corp., Box 746, Baltimore, Md.

Published by The Institute of Radio Engineers, Inc., for the Professional Group on Automatic Control, 1 East 79 Street, New York 21, N. Y. Responsibility for the contents rests upon the authors, and not upon the IRE, the Group or its members. Individual copies of this issue and all available back issues may be purchased at the following prices: IRE members (one copy) $2.25; libraries and colleges $3.25; all others $4.50. Annual subscription price: libraries and colleges, $12.75; non-members $17.00.

This is the first issue of the PGAC TRANSACTIONS *to have papers selected and arranged by a Guest Editor. For this work, we gratefully thank Dr. Robert Kramer of the M.I.T. Electronic Systems Laboratory. In addition, this is the first regular issue of PGAC* TRANSACTIONS *to feature conference papers which have been subjected to a full review procedure. We hope that this policy can be continued and that it will meet with the approval of our members and others interested in automatic control.—The Editor*

Foreword

The 1961 Joint Automatic Control Conference is the second conference on automatic control sponsored jointly by the AIChE, AIEE, ASME, IRE, and the ISA. This joint action by the five societies is an attempt to reduce the number of meetings on control systems and at the same time, hopefully, to improve on the quality of the papers. Judging from the interest generated at the first meeting, the idea of a joint conference is meeting with general approval. As the mechanism for holding these meetings is refined with experience, their success should be even greater.

At present, the publication of the papers presented at these conferences is the responsibility of the individual societies since, as yet, no combined proceedings has been agreed upon. For the first conference (1960), the IRE-PGAC papers were reviewed and selected by a special program committee independent of the regular TRANSACTIONS review board. The large number of papers and the short time available for their review precluded any participation by the regular TRANSACTIONS review board. Since the selected papers were thus not of assured TRANSACTIONS quality, they were published in a special JACC Record issue of the TRANSACTIONS in preprint format.

This year, however, the procedure was modified in two important aspects. To ensure that the reviews of these JACC papers be up to the standards of the regular TRANSACTIONS papers, one of the three reviews given each of the papers submitted was by the regular PGAC TRANSACTIONS review board, under the direction of the regular PGAC *Editor*. The IRE Program Committee arranged to have two additional reviews of each paper. The selection of papers was made on the basis of all these reviews. In this way, the high standards of the TRANSACTIONS papers were injected into the selection of Conference papers, even though many new reviewers were pressed into service.

In addition to this review change and partly because of it, decision was made to publish this year's papers in a regular issue of the TRANSACTIONS. This, in effect, accords these conference papers a status equal to those of the regular TRANSACTIONS, with fine letter-press printing and incorporating reviewers' comments. Furthermore, an attempt was made to have this issue ready for distribution at the Conference. Only you, the reader, will know the degree of our success in having our papers published concurrent with the Conference.

Apropos of our attempt to gauge the success of a joint Conference, one good indication of the popularity of this Conference is the great amount of material submitted for our consideration. We received no less than 31 papers, despite the fact that there were only three months between the initial call for papers and the deadline for submitting them. It was quite obvious that program limitations would preclude our presenting a great many of them, and, therefore, a considerable number had to be eliminated. This presented the Program Committee with a problem and a prize. The problem was reviewing this large number of papers carefully and selecting the best; the prize, of course, is a potentially excellent group of papers. We finally accepted 16 of the 31 submitted—an acceptance of 52 per cent—and we feel that the IRE-PGAC group of papers is of high quality. We sincerely hope that, rather than discouraging authors from submitting papers to the IRE because of the high rejection rate, this will encourage the submission of high-quality papers to a Conference of recognized high quality.

In conclusion, we should like to express our personal thanks to all those who did the real work of reviewing and evaluating the articles which were submitted.

—ROBERT KRAMER, *Guest Editor*

The Issue in Brief

On a Property of Optimal Controllers with Boundedness Constraints, **H. L. Groginsky**

This paper deals with a theory for optimal control systems designed to operate a plant of known characteristics. It is assumed that only limited changes in the system characteristics can be effected by the control variables at the designer's disposal.

It is shown that within the assumed limitations for a wide class of inputs and systems and for a certain class of measures of the system performance, an optimal system behaves as a relay or switched system during the transient period and as a continuous system during periods in which the input is reproduced identically. A procedure is described for determining the switching times during the transient period in terms of the permissible measurements.

A Minimal Time Discrete System, **C. A. Desoer and J. Wing**

A sampled-data control system with a plant having real poles and limited input is considered. The plant forcing function which will bring the system to equilibrium in a minimum number of sampling periods is desired; this function is called an optimal strategy.

To implement this particular optimal strategy, we define a surface in state space called the critical surface. It is shown that this optimal strategy will be generated by the following procedure: at the beginning of each sampling period, the distance ϕ from the state of the system to the critical surface is measured along a fixed specified direction; if $\phi \geq 1$ (or ≤ -1), then the forcing function for that sampling period is $+1$ (or -1); if $|\phi| < 1$, then the forcing function is ϕ. For a third-order plant, it is shown that the critical surface has certain properties which lead to a simple analog computer simulation.

Theory and Design of High-Order Bang-Bang Control Systems, **M. Athanassiades and O. J. M. Smith**

A complete analysis and design is presented of a nonlinear controller to minimize the response time of a limited input plant whose transfer function has N real roots.

The design procedure is based exclusively on the concept of the switching hypersurface of the system in N-dimensional state space and on the concept of the "distance function" from the state point to the switching hypersurface.

The linear and nonlinear transformations performed by a nonlinear computer upon the error and its time derivatives, in order to generate the optimal amplitude limited input to the plant, are described in detail, and properties of the switching surfaces, which are subsets of the switching hypersurface, are also described.

Model Feedback Applied to Flexible Booster Control, **G. E. Tutt and W. K. Waymeyer**

The fundamental control problems in the design of flexible space vehicle boosters center around the control of an aerodynamically unstable airframe in a dynamic wind shear (jet stream) environment. This paper presents a feedback model approach to this design problem which shows promise. This system does not adapt to body bending, but instead is contrived to ignore it.

A conventional attitude control system is assumed in which rigid body control considerations have been used to design the control loops. A model of the plant (airframe rigid body dynamics) is derived. The attitude and rate feedbacks are now syntheszied by a combination of actual and model rate and attitude. Bending and similar dynamic effects are filtered from the actual attitude and rate signals as required, and the information thus rejected is supplied from the model of the plant. It is readily shown that the performance of this system in response to commands is substantially identical in all important respects to the original rigid body system.

Terminal Control System Applications, **E. A. O'Hern and R. K. Smyth**

This paper deals with certain theoretical extensions of earlier work on terminal control techniques and their application to an aircraft landing system. The case considered is of a system having a second-order response from altitude rate command to altitude rate. The terminal time equations for this case are developed together with the various closed-loop weighting functions by transform methods. From the terminal equations developed, the terminal controller equations are synthesized for a two-condition terminal controller which controls altitude and altitude rate at the terminal (touchdown) time. The mechanization of these terminal controller equations are presented.

The flight test results of the terminal control system using the system developed in this paper are described briefly. A comparison between flight test and simulation results is included.

A Parameter-Perturbation Adaptive Control System, **R. J. McGrath, V. Rajaraman, and V. C. Rideout**

Theoretical and simulator studies of some new forms of a parameter perturbation self-adaptive system are presented in this paper. Here, the response of the control system is subtracted from that of a reference model to obtain the error signal. As in the previous studies, the controllable parameters of the system are sinusoidally perturbed. Somewhat different methods of processing the perturbation and error signals have been employed to obtain the parameter control signal which is used in the automatic optimization of the control system.

Computer studies are made with both random and deterministic input signals and parameter disturbances. The results are compared with the analytical results.

Transfer-Function Tracking and Adaptive Control Systems, **C. N. Weygandt and N. N. Puri**

In an adaptive system in which the plant parameters are varying, it is necessary to track or measure the plant parameters. Two separate schemes are proposed for tracking the transfer function of a multi-order system.

The first scheme is based on perturbing the normal process with small amplitude sinusoidal perturbation signal consisting of different frequencies. The tracking system is a closed-loop system.

The second scheme does not depend upon perturbation signal, but does require knowledge of the form of the transfer function of the system. It is more suitable for tracking systems which have a number of first-order lag units connected in tandem.

Adaptive Servo Tracking, **A. I. Talkin**

This paper describes a self-adapted sampled-data radar tracking loop. The tracking loop may be considered to be a low-pass filter with a variable bandwidth. The loop is designed to adapt rapidly to changes in the input signal by monitoring both the apparent error and the loop output.

Results show a mean tracking error 25–34 per cent less than that of a comparable linear system, at a receiver SNR of 10 db.

Precision of Impulse-Response Identification Based on Short, Normal Operating Records, **R. B. Kerr and W. H. Surber, Jr.**

An identification scheme is presented for estimating in a short time the impulse response of a system from normal operating records.

Maximum-likelihood estimates of the impulse response are discussed, and a "sufficiency" criterion on the input signal is defined based upon the expected integrated squared difference between the actual and estimated impulse responses. Examples of sufficient and insufficient test signals are given in terms of a "sufficient record length" criterion. Experimental results illustrating this latter criterion are presented.

The time variation of the system parameters sets an upper bound on the useful record length. Some preliminary results relating this time variation to the useful record length are also presented.

A Technique of Linear System Identification Using Correlating Filters, **W. W. Lichtenberger**

Random signals and cross correlation can be used to determine the impulse response of a system. If, instead of a random testing signal, one is employed which has certain properties similar to the random signal, a linear filter may be constructed which performs the necessary cross correlation. No multiplication is involved, and the output is a continuous function of time. Thus, a single filter obtains the impulse response for all values of time. A disadvantage of this method is that in a practical situation, there is usually a great amount of noise present. This "noise" consists mostly of process actuating signals since the measurements must be made "on line."

A Modified Lyapunov Method for Nonlinear Stability Analysis, **D. R. Ingwerson**

The Lyapunov stability criterion deals with arbitrarily small disturbances. A generalization of the original theorem which applies to arbitrarily large and arbitrarily small disturbances, and to intermediate conditions as well, is given in this paper.

In contrast to the success that has been achieved in advancing the theoretical concepts of stability by this method, little has been accomplished in the way of formulating practical means for applying them to specific problems. A method which is easily applied to many of the systems encountered in automatic control and which has given good results for numerous examples is presented here.

A Mean-Weighted Square-Error Criterion for Optimum Filtering of Nonstationary Random Processes, **G. J. Murphy and K. Sahara**

A procedure for use in the design of a physically realizable time-invariant linear system for optimum filtering of a nonstationary random process in the presence of nonstationary random noise is presented. The criterion used to measure system performance is a mean-weighted square error.

It is shown that the use of this generalization of the mean square-error criterion leads to a generalization of the Wiener-Hopf integral equation. A technique for solving this modified Wiener-Hopf integral equation is presented, and the application of the theory is illustrated in an example of the optimum synthesis of a missile interception system.

A General Performance Index for Analytical Design of Control Systems, **Z. V. Rekasius**

A performance index which enables one to specify the desired response of the optimum system in terms of the differential equation describing the response of an ideal model is proposed.

A simple straightforward procedure of calculating this performance index is outlined in the paper. This procedure consists of solution of a set of linear algebraic equations. Simultaneous solution of these algebraic equations yields the value of performance index in terms of gain and time constants of the actual system. It is then a simple matter to calculate the numerical values of the free gain and time constant parameters for the optimum system (*i.e.*, to minimize the performance index). The procedure of optimization is illustrated by means of an example of a third-order system.

Stability of Servomechanisms with Friction and Stiction in the Output Element, **P. K. Bohacek and F. B. Tuteur**

Servomechanisms with friction in the output element are often observed to oscillate, even though the Bode diagram indicates stability. This paper investigates the conditions for this instability and the type of oscillation that can occur. It finds that an overdamped system with a lag equalizer is stable if $L<2C/C-1$, where L is the lag ratio and $C=$static friction$+$Coulomb friction; with a lag-lead equalizer it is stable if

$$\frac{L}{1+a/b}<\frac{2C}{C-1},$$

where a/b is the ratio of the two zeros of the network. Experimental results that correlate with the theory are also included.

Sensitivity Considerations for Time-Varying Sampled-Data Feedback Systems, **J. B. Cruz, Jr.**

A synthesis procedure for linear time-varying sampled-data feedback systems is described. The plant and compensators are characterized by transmission matrices first introduced by Friedland. Part of the specification involves a deviation or error matrix for the time-varying plant and an allowable deviation matrix for the closed-loop system. Noise considerations are also included. Using a technique analogous to that of Horowitz which was originally used for fixed multivariable continuous control systems, the digital compensator transmission matrices are derived. The corresponding time-varying digital compensators are realized by means of zero-order hold circuits, switches, resistors, and adders. An example using a digital computer simulation is included.

Direct Cycle Nuclear Power Plant Stability Analysis, **D. Buden and R. F. Miller**

A power plant with a heat exchanger such as a nuclear reactor substituted for the conventional chemical interburners in a jet engine will cause a considerable change in dynamic performance. The instantaneous power generated by the heat source is not the same as the instantaneous power delivered to the turbine. The basic control problems are analyzed using fixed control parameters and partial derivatives around a given operating point. A mathematical criterion is developed and correlated with power plant test data.

An understanding of the inherent limitations of combining a reactor, or any heat exchanger having a thermal lag, with a basic jet engine makes it possible to devise a means of control. The introduction of an effective operational speed control makes it possible to operate a complete power plant under any desired condition.

On a Property of Optimal Controllers with Boundedness Constraints*

HERBERT L. GROGINSKY†, MEMBER, IRE

Summary—This paper deals with a theory for optimal control systems designed to operate a plant of known characteristics. It is assumed that error-free measurement of the system input and the system behavior is permitted, but only limited changes in the system characteristics can be effected by the control variables at the designer's disposal.

It is shown that within this limitation for a wide class of inputs and systems, and for a certain class of measures of the system performance, an optimal system behaves as a relay or switched system during the transient period, and as a continuous system during periods in which the input is produced identically.

A procedure is described for determining the switching times during the transient period in terms of the permissible measurements. The result is the design of the optimal controller. Typically, its realization requires analog computation of the switching function and digital switching of the control variables.

The design of a second-order regulator system, in which the control variable is the gain in the feedback path, is obtained. Marked improvement in the system performance is noted.

I. Introduction

THE problem discussed in this paper is the following:

1) given a *basic system* whose characteristics are completely known (its set of differential equations including parameters, adjustable parameters, and available forcing functions are specified),
2) given any desired noise-free measurement up to the present of the system input and the performance of the basic system,
3) given a means of changing the performance of the basic system while it is in operation (note that only a limited amount of such control is available),

how then can a controller be designed which automatically and continuously selects the values of the adjustable parameters within the admissible range which optimizes the over-all system performance on a wide class of inputs.

In terms of a mathematical model, the problem is to find a set of functions $Y[\mathbf{x}(t), T]$ in which $\mathbf{x}$ is the total system state of the system (*i.e.*, the internal state of the basic system plus the state variables describing the input) governed by[1]

$$\dot{\mathbf{x}}(\tau) = \mathbf{g}[\mathbf{x}(\tau), Y(\tau)]$$
$$\mathbf{x}(t) = \mathbf{x}, \tag{1}$$

which minimizes

$$J = \int_t^{t+T} G(\epsilon(\tau))d\tau \tag{2}$$

where

$$\epsilon(t) = \mathbf{a}'\mathbf{x}$$
$$\mathbf{a} \triangleq \text{constant vector}$$

under the constraint that

$$|\, y_{ij}(\mathbf{x}, t)\,| \leq k_{ij} = \text{fixed constant.} \tag{3}$$

It is clear that the knowledge of Y as a function of the measurable quantities $\mathbf{x}$ is sufficient to specify completely the design of the optimal controller. It is also clear from the formulation of the problem that only the present state of the system determines the present adjustment of the control variable so that storage of past state values is unnecessary.

In this analysis, only the special case in which the adjustable parameters of the basic system enter the system (1) linearly is considered. In other words, although the basic system may contain nonlinearities, the system equations are to have the form

$$\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}) + Y\mathbf{h}(\dot{\mathbf{x}}). \tag{4}$$

The generic form of the systems under consideration is shown in Fig. 1.

II. Properties of an Optimal Controller

A. Summary

The object of this section is to establish conditions under which an optimal adaptive controller sets the control variables to their extreme allowable values. This type of control function is called the *switching mode*. The primary function of the controller during such operations consists of selecting the times at which the control variables are switched from one extreme to the other.

Criteria of the form (2), in which $G(\epsilon)$ is continuous and differentiable and monotonic in $|\epsilon|$, and inputs which are noisefree and bounded lead to the need for the switching mode when the system is in transient operation. The switching mode of operation terminates when the input can be reproduced identically, with the control parameter within its prescribed bounds. However, if the system has entered its continuous mode, but the input cannot be reproduced identically throughout the entire interval of concern, the system must eventually revert to the switching mode. It is shown that the continuous mode terminates when the system input can no

* Received by the PGAC, November 14, 1961; revised manuscript received, March 10, 1961. The research was supported by the Electronics Research Directorate of the AF Cambridge Research Center.

† Raytheon Co., Advanced Dev. Lab., Wayland, Mass. Formerly at Electronics Res. Labs., New York, N. Y.

[1] Unless otherwise noted, lower case letters denote scalars, boldface letters denote vectors and upper case letters denote matrices.

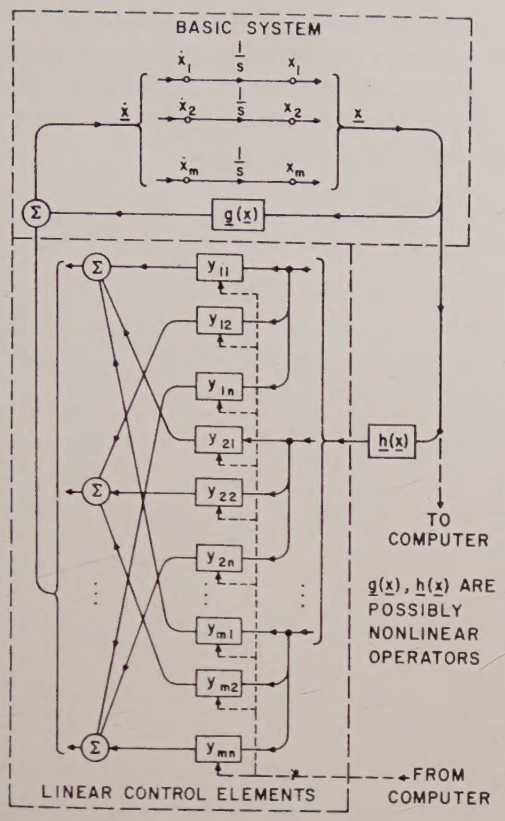


Fig. 1—General form of linearly controlled basic system.

longer be reproduced identically. Prior termination in anticipation of the switching mode cannot improve the performance of an optimal system.

The simplest, most straightforward derivation of this result is obtained through the use of dynamic programming. However, because it is necessary to assume differentiability of the minimal loss function, a fact which is not known *a priori*, the derivation lacks rigor. The Euler-Lagrange method provides the rigorous treatment, and in addition resolves the problem of the onset of the switching mode.

The solution obtained by dynamic programming is important because it provides a computational algorithm with which the design of the optimal controller can be obtained.

B. A Theorem on Optimal Adaptive Controllers

Theorem:

1) Let $y(x, t)$ be a set of m control functions for a basic system which is linearly controlled, *i.e.*, the system governed by the set of equations

$$\dot{x} = g(x) + H(x)y; \qquad x(0) = x_0 \tag{5}$$

where $g(x)$, $H(x)$ are sets of functions of class C^1 in x.[2]

[2] A function of n variables $f(x)$ defined on an open region R belongs to class C^k in R if its first k partial derivatives exist and are continuous at every point x in R.

2) Let $\epsilon(t) \triangleq$ system error and the first p derivatives of ϵ be such that

$$\begin{aligned} \epsilon &= \epsilon_0(x) \\ \dot{\epsilon} &= \epsilon_1(x) \\ &\vdots \\ \epsilon^{(p)} &= \epsilon_p(x). \end{aligned} \tag{6}$$

3) Let $G(\epsilon, \dot{\epsilon}, \cdots, \epsilon^{(p)})$ be a function of class C^1 jointly in all its variables, having the properties that:

a) $G(\epsilon, \dot{\epsilon}, \cdots, \epsilon^{(p)}) = 0$ if and only if

$$\epsilon = \dot{\epsilon} = \cdots = \epsilon^{(p)} = 0,$$

b) $$\frac{\partial G}{\partial \epsilon^{(j)}} \begin{matrix} > 0; & \epsilon^{(j)} > 0 \\ < 0; & \epsilon^{(j)} < 0 \end{matrix} \qquad j = 0, 1, \cdots, p.$$

4) Let $\sigma(x) = G(\epsilon_0(x), \epsilon_1(x), \cdots, \epsilon_p(x))$ be such that

$$\sigma_x H(x) = 0 \tag{7}$$

where

$$\sigma_x = \begin{bmatrix} \dfrac{\partial \sigma}{\partial x_1} \\ \vdots \\ \dfrac{\partial \sigma}{\partial x_n} \end{bmatrix}.$$

5) Let $y_{\text{opt}}(x, T)$ be a control function such that

$$J(x, T; \{y_{\text{opt}}\}) = \min_{\{y\}} \int_0^T G(\epsilon, \dot{\epsilon}, \cdots, \epsilon^{(p)})dt \tag{8}$$

under the additional constraint that each member of the set is bounded, *i.e.*,

$$|y_{j\,\text{opt}}| \leq k_j \qquad j = 1, 2, \cdots, m. \tag{9}$$

6) Let at least one among $\epsilon, \dot{\epsilon}, \cdots, \epsilon^{(p)}$ be different from zero at some time t on the interval $0 \leq t \leq T$. *Then at that time t the function* $y_{\text{opt}}(t)$ *has the property that*

$$|y_{j\,\text{opt}}| = k_j \qquad j = 1, 2, \cdots, m. \tag{10}$$

Before demonstrating the proof of this theorem a brief discussion of the meaning of the assumptions 1) through 6) is given. Some of these will be further amplified in the derivations which follow.

1) The system equation (4), corresponding to the system shown in Fig. 1, is first rewritten in a form which explicitly separates all of the *independent* control parameters. A general system may have a single parameter governing the interactions on a number of states. In other words, various members of Y in (4) may be equal, $y_{ij} = y_{kl}$. Eq. (5) shows only the m variables which can be varied by the controller; all other redundant variables are eliminated. In general there are far fewer control variables than the number of system states.

2) The system error and its first p derivatives are assumed to be defined strictly by the state variables, and none of these involve any control parameter y_j.
3) The criterion function is a distance-measuring function in the sense that the value of the function is positive when the system error or any of its first p derivatives differ from the desired condition, namely, $\epsilon_j = 0$, $j = 0, 1, \cdots, p$, and further, the value increases as the deviation of any one of the ϵ_j's from zero increases.
4) The control function is assumed to have no immediate effect on the system error or any of its p derivatives. This condition is an alternate expression of the information in assumption 4).
5) The function $\boldsymbol{y}_{\text{opt}}$ is assumed to be an optimal control function satisfying boundedness constraints. The fact that the bounds are symmetric about zero is unimportant.
6) At time t the system is assumed to be in a transient condition.

Conclusion: At the time t all parameters must be set to their extreme values.

A credible proof of this theorem is offered first by the method of dynamic programming.

Let

$$f(\mathbf{x}, T) = \min_{\{y\}} J(\mathbf{x}, T; \{\boldsymbol{y}\}) \tag{11}$$

with

$$|y_j| \leq k_j \qquad j = 1, \cdots, m$$

where J, $\mathbf{x}$ are as defined previously. It is important to note that f is independent of $\boldsymbol{y}$ since, assuming a solution exists, given $\mathbf{x}$ and T the optimal control function $\boldsymbol{y}_{\text{opt}}$ is determined on the entire interval $0 \leq t \leq T$.

The technique of dynamic programming is used to determine a functional equation for f as follows. Writing

$$\begin{aligned}\int_0^T \sigma(\mathbf{x})dt &= \int_0^T G[\epsilon_0, \epsilon_1, \cdots, \epsilon_p]dt \\ &= \int_0^\Delta \sigma(\mathbf{x})dt + \int_\Delta^T \sigma(\mathbf{x})dt \\ &= \int_0^\Delta \sigma(\mathbf{x})dt + \int_0^{T-\Delta} \sigma[\mathbf{x}(\tau + \Delta)]d\tau\end{aligned}$$

and employing the principle of optimality,[3] the functional equation

$$f(\mathbf{x}, T) = \left\{\min_{y'}\right\} \left\{\int_0^\Delta \sigma(\mathbf{x})dt + f(\mathbf{x}(\Delta), T - \Delta)\right\}$$

is obtained. In this equation $\{\boldsymbol{y}'\}$ represents the set of control functions on the initial interval $0 \leq t \leq \Delta$.

[3] R. Bellman, "Dynamic Programming," Princeton University Press, Princeton, N. J., 1957. See especially chap. 9, pp. 245–267.

If the input $r(t)$, now considered to be one of the system states, has a *piecewise continuus* derivative,[4] and if the admissible control functions $\boldsymbol{y}$ likewise have piecewise continuous derivatives then, first, the system trajectory $\mathbf{x}(t)$ is continuous,[5] and second,

$$\int_0^\Delta \sigma(\mathbf{x})dt \xrightarrow[\Delta \to 0]{} \sigma(\mathbf{x}_0)\Delta + o(\Delta), \tag{12}$$

except possibly on a set of measure zero.

Likewise, *if the first partial derivatives of f with respect to each of its variables are continuous*, then $f(\mathbf{x}(\Delta), T - \Delta)$ has the MacLaurin series expansion

$$\begin{aligned}&f(\mathbf{x}(\Delta), T - \Delta) \\ &\quad = f(\mathbf{x}_0, T) + \left(f_x' \frac{d\mathbf{x}(\Delta)}{d\Delta} - f_T\right)\Delta + o(\Delta)\end{aligned} \tag{13}$$

where

$$f_T = \frac{\partial f}{\partial T}$$

and f_x is the vector

$$f_x = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \cdot \\ \cdot \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix}$$

$$\begin{aligned}n = n_s + n_i = &\text{ number of system states} \\ &+ \text{number of input states}\end{aligned}$$

and $'$ denotes the transposition operation.

Under these assumptions the remainder terms are of order greater than Δ as indicated.[6] Since under the previous assumptions the state variables have piecewise continuous derivatives, the derivatives in (13) can be replaced by their values at time $t = 0$, as given by the system equation (5). This permits (13) to be written as

$$\begin{aligned}&f(\mathbf{x}(\Delta), T - \Delta) \\ &\quad = f(\mathbf{x}_0, T) + \Delta\{f_x'(\boldsymbol{g}(\mathbf{x}_0) + H(\mathbf{x}_0)\boldsymbol{y}_0) - f_T\} + o(\Delta),\end{aligned} \tag{14}$$

where $\boldsymbol{y}_0$ denotes the values of the control parameters at time t.

Combining (12) and (14) and passing to the limit as Δ tends to zero, the nonlinear partial differential equation

$$0 = \min_{y} \{\sigma(\mathbf{x}) + f_x'[\boldsymbol{g}(\mathbf{x}) + H(\mathbf{x})\boldsymbol{y}] - f_T\} \tag{15}$$

[4] The function $r(t)$ is said to be piecewise differentiable on the interval $0 \leq t \leq T$ if it is differentiable everywhere except for a set of measure zero, where it is permitted to have only finite discontinuities.

[5] E. A. Coddington and N. Levinson, "Theory of Ordinary Differential Equations," McGraw-Hill Book Co., Inc., New York, N. Y., chap. 1, p. 3; 1955.

[6] P. Franklin, "Treatise on Advanced Calculus," John Wiley and Sons, Inc., New York N. Y., pp. 137–138; 1940.

is obtained. Note that whereas in (11), $\{y\}$ is a class of admissible control functions, *in* (15) y *is simply the set of values of the control parameters at time* $t=0$. Note also that the knowledge of y at time $t=0$ as a function of the initial state x_0 and the interval length T gives all the information needed to specify y as a function of x and $T-t$ throughout the entire interval. Hence, (15) has been written for an arbitrary time t in the interval, which means that y_0 has been replaced by y, x_0 by x, and T by $T-t$.

Because f is the *minimal* loss function, it is *not* a function of y, hence, the only terms in (15) which depend on y are linear in y. Since the values of y are bounded by k, the minimum in (15) is obtained by setting each of the variables to

$$y_{j\,\mathrm{opt}} = -k_j \operatorname{sgn} \sum_{i=1}^{n} f_{x_i} h_{ij}(x) \qquad j = 1, \cdots, m; \quad (16)^7$$

that is, the values of y lie at the bounds except when

$$W_j(x, T-t) = -\sum_{i=1}^{n} f_{x_i}(x, T-t) h_{ij}(x) = 0. \quad (17)$$

The set of functions $W_j(x, T-t)$ henceforth are called the *switching functions*. Geometrically, (17) can be said to describe the boundaries of the regions in the *phase space* in which the sign of the jth control variable is constant.

Now if at any time $W_j(x, T-t)=0$, *i.e.*, a boundary surface is reached, but any of the remaining terms in (15) are not zero, (15) remains in force as the determining equation for the system and $y_{j\,\mathrm{opt}}$ can be selected arbitrarily. In particular, if $\sigma(x)\neq 0$ which amounts to $G(\epsilon, \dot{\epsilon}, \cdots, \epsilon^{(p)})\neq 0$, the magnitude of the control parameters can be set to their respective bounds. Assumption 2) states that $G(\epsilon, \dot{\epsilon}, \cdots, \epsilon^{(p)})=0$ implies that $\epsilon=\dot{\epsilon}=\cdots=\epsilon^{(p)}=0$, hence, the result of the theorem is demonstrated.

The result can be stated alternatively as follows:

As long as the system output differs from the system input (*i.e.*, the system is in the transient condition, namely at least one among ϵ, $\dot{\epsilon}$, $\ddot{\epsilon}$, $\cdots$ is different from zero), *an optimal system of the form* (5) *uses relay or switched type control.*

The transient period terminates whenever the system error and all its derivatives are zero. This corresponds to the case in which all terms of (15) are zero. A sufficient condition for this to occur is that the system reproduce the input identically at time t, and furthermore reproduce it identically throughout the remainder of the interval $T-t$, with the amount of control available, *i.e.*, $y_j \leq k_j$, $j=1, \cdots, m$. At such a time the value of the control variable y_j may not be at any extreme, nor is it arbitrary. In fact, the values to which y must be set at this time can be computed by setting the left-hand side of (6) to zero, and then solving this system of equations together with the system equation (5) simultaneously for the instantaneous parameter values. This type of operation henceforth is referred to as the *continuous mode.*

It is to be noted that the demonstration given above is deficient because the assumption of the existence of a set of continuous first partial derivatives of the minimal loss function f cannot be justified *a priori* on the basis of the hypotheses of the theorem. Nonetheless the method is extremely useful, primarily because it provides a practical computational algorithm for the evaluation of the switching functions through (16), as is described in subsequent sections. Fortunately, the conventional Euler-Lagrange technique can be employed to rigorously justify the conclusion of the theorem as is demonstrated in Appendix I. Indeed the latter is also useful in resolving the issue of when the switching mode begins when the system has previously been operating in the continuous mode.

The point in question pertaining to the onset of the switching mode can be stated as follows. Suppose that an input to the system has been applied which can be reproduced identically on an initial interval $0 \leq t \leq t_1$ with $|y_j| \leq k_j$, $j=1, \cdots, m$ but to do so on the interval $t_1 \leq t \leq T$ requires at least one of $|y_j|$ to exceed its bound k_j. It is clear that in the interval $t_1 \leq t \leq T$ the controller operates in the switching mode, but it may be that the optimal controller would begin this mode of operation prior to t_1 in anticipation of the desired response. In other words, the question is whether there are any circumstances under which it would prove profitable to deliberately introduce an error into the system at a time when the input can be reproduced identically in order to reduce errors which will occur at a future time. The general theorem stated above denies that this should be done. In Appendix I the proof of the theorem includes a detailed study of this condition, and it is shown that *an optimal system does not begin switching prior to* t_1.

C. A Second-Order System with a Single Variable Parameter

For the sake of explicitness a simple prototype system of the kind shown in Fig. 2(a) is considered. The variable parameter is taken to be the gain in the feedback loop, so that the block diagram of the system is that shown in Fig. 2(b).

The differential equations of this system are

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\alpha x_2 - \beta y x_1 + \beta r, \end{aligned} \quad (18)$$

where y designates the controllable parameter and the bound on y is $|y| \leq 1$. That is, the feedback can be varied from negative to positive feedback.

Consider that the loss function is of the type defined

[7] $\operatorname{sgn} z = \begin{cases} z/|z|; & z \neq 0 \\ 0; & z = 0 \end{cases}$.

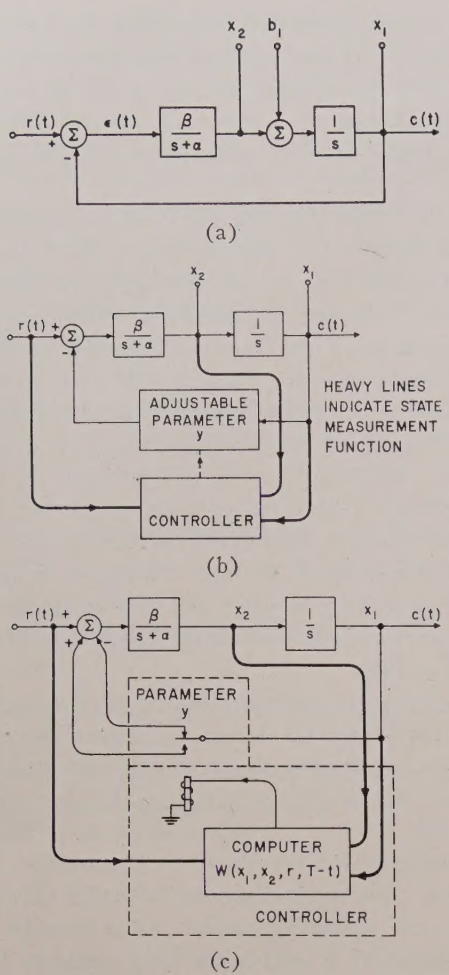


Fig. 2—(a) Typical second-order system. (b) Second-order system with variable feedback gain y. (c) Implementation of an optimum second-order system.

in (2). Then let

$$f(x_1, x_2, t, T) = \min_{\{y\}} \int_0^T G[\epsilon(t+\tau)]d\tau \tag{19}$$

where

$$\epsilon(t) = x_1(t) - r(t). \tag{20}$$

Checking the conditions of the theorem, it is seen that hypothesis 1) is satisfied by the system equation (18), hypothesis 2) is satisfied by the definition (20) and hypothesis 3) is satisfied by choosing an appropriate G. Hypothesis 4) is checked as follows.

$$\sigma(\mathbf{x}) = G(x_1 - r)$$

$$\sigma_x = \frac{dG(\epsilon)}{d\epsilon}\begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{21}$$

In (21) the input state has been ignored because the parameter y in no way affects the transmission of the input into the system. Thus,

$$\sigma_x'\boldsymbol{h}(\mathbf{x}) = \frac{dG(\epsilon)}{d\epsilon}[1\ 0]\begin{bmatrix} 0 \\ -\beta x_1 \end{bmatrix} \equiv 0,$$

which is the assertion of hypothesis 4).

Thus it is concluded that when $\epsilon \neq 0$

$$\begin{aligned} y_{\text{opt}} &= -\operatorname{sgn} f_x'\boldsymbol{h}(\mathbf{x}) \\ &= \operatorname{sgn} W(x_1, x_1, r, T-t) \\ &= \operatorname{sgn} x_1 f_{x_2} \end{aligned} \tag{22}$$

derived from the nonlinear partial differential equation

$$0 = \min_y \{G(\epsilon) + x_2 f_{x_1} + (\beta r - \alpha x_2 - \beta y x_1) f_{x_2} - f_T\}.$$

In other words, an optimal controller for the system shown in Fig. 2(b) would generally use a relay transfer contact for the feedback gain element with a computer controlling the sense of the relay switching as shown on Fig. 2(c).

III. Example of a Design of a Second-Order Regulator System

A. Summary

In this section, the basic system described in Section II-C and designed as an autonomous system previously[8] is now designed for step inputs. The details of the calculations are much more involved even in this elementary case. Because of the location of the control element in the basic system, the results of the autonomous case are not readily extended to this new condition. The design of a particular system is accomplished with the aid of a digital computer. The results of the computer study show that the optimal switching functions tend asymptotically to the straight line segments in the phase plane obtained previously. The realization of the system is described. The design procedure for an nth order system is also described.

B. Description of the System

The system to be designed in this section has been described in Section II-C. This system is to be used as a regulator, hence the input is a constant r_0. The system is to respond optimally to disturbances from its equilibrium condition, which is the reproduction of r_0 identically. Equivalently the disturbances can be attributed to step changes in the input, so that the design is also appropriate for a system designed to follow discrete displacements.

The criterion function is the integrated square error. In this case it is possible for the basic system to reproduce the steady input identically, as can be seen by examining the steady state behavior of the basic system with $y=1$. In fact if $y=1$, the infinite integral, (19) with $T=\infty$ exists, which shows that there is a bound for the infinite integral. Thus a solution to the optimization problem exists.

The problem posed in this section does not reduce directly to the autonomous case by translation of the coordinate in the phase space because of the location of the nonlinear control. To show this, consider the effect of translating x_1 by r_0. Let

[8] H. L. Groginsky, "On the Design of Adaptive Systems," 1958 IRE National Convention Record, pt. 4, pp. 161–167.

$$x_1 = x_3 + r_0$$
$$x_2 = x_2. \tag{23}$$

Then substituting in (18), one obtains

$$\dot{x}_3 = x_2$$
$$\dot{x}_2 = -\alpha x_2 - \beta y x_3 + \beta(y-1)r_0. \tag{24}$$

But this new system is also not autonomous as is readily observed. More generally, there is no pair of functions $f_1(r_0, \alpha, \beta), f_2(r_0, \alpha, \beta)$ for which the substitutions

$$x_1 = U_1 + f_1$$
$$x_2 = U_2 + f_2$$

reduce (24) to a form independent of r_0.

C. The Switching Function

The rigorous procedure for evaluating the switching function is begun by evaluating the loss function for finite but small T under each choice of the control variable. These loss functions are then used to determine the switching function corresponding to this interval T through (19).

In this case, as in the autonomous case, since the input can be reproduced identically in the steady state if $y \equiv 1$ at all times, the solution when $T = \infty$ can be arrived at immediately. This is done by examining the region in the phase space in which $y = 1$ for all time, and (19) is $+1$ at all points on the system trajectory.

Since the system is linear and autonomous with $y = 1$, the loss function and the switching function for this case can be obtained immediately from the autonomous case. The result is that

$${}^{+}f^0(x_1, x_2) = \frac{(x_1 - r_0)^2}{2\alpha} + \frac{[\alpha(x_1 - r_0) + x_2]^2}{2\alpha\beta} \tag{25}$$

and

$${}^{+}f_{x_2}{}^0(x_1, x_2) = \frac{x_2 + \alpha(x_1 - r_0)}{\alpha\beta}. \tag{26}$$

The region $R_+{}^0$, in which the control variable is $+1$ and no further changes in the switching function is required, is

$$R_+{}^0 = \left\{ x_1, x_2: \begin{array}{c} x_2 + \alpha(x_1 - r_0) > 0; \\ x_1 > 0; \\ x_2 + \rho_2(x_1 - r_0) > 0 \end{array} \right\}. \tag{27}$$

where

$$\rho_1 = \frac{\alpha + \sqrt{\alpha^2 - 4\beta}}{2}, \qquad \rho_2 = \frac{\alpha - \sqrt{\alpha^2 - 4\beta}}{2}$$

and is shown in Fig. 3(a) (next page). The latter line is obtained from the condition that the trajectory remain in the $R_+{}^0$ region for all time.

The region $R_-{}^1$ adjoining $R_+{}^0$, in which $y = -1$, is found as follows. The points in the phase space not in $R_+{}^0$ are considered. Among these, the set is selected which, if taken as the initial disturbances of the system, leads to trajectories terminating on the boundaries of $R_+{}^0$ when y is fixed at -1. These points are the ${}^{-}R^1$ region

$${}^{-}R^1 = \{x_1, x_2: \text{ if } y = -1 x_1(T),\ x_2(T) \in R_+{}^0\}$$
$$\text{for some } T < \infty. \tag{28}$$

Now if $y = -1$

$$x_1(t) = Ce^{-\rho_3 t} + De^{\rho_4 t} + r_0 \tag{29a}$$
$$x_2(t) = -\rho_3 Ce^{-\rho_3 t} + \rho_4 + De^{\rho_4 t}, \tag{29b}$$

where

$$C = \frac{\rho_4(x_1 + r_0) - x_2}{\rho_3 + \rho_4} \tag{29c}$$

$$D = \frac{\rho_3(x_1 + r_0) + x_2}{\rho_3 + \rho_4} \tag{29d}$$

and

$$\rho_3 = \frac{\sqrt{\alpha^2 + 4\beta} + \alpha}{2} \qquad \rho_4 = \frac{\sqrt{\alpha^2 + 4\beta} - \alpha}{2}.$$

Examination of these equations shows that the potential ${}^{-}R^1$ region is the shaded area shown in Fig. 3(a).

Now for any initial point in the potential ${}^{-}R^1$ region, the loss function is

$${}^{-}f^1(x_1, x_2) = \int_{\phi}^{T_1} [r_0 - x_1(t)]^2 dt + {}^{+}f^0[x_1(T_1), x_2(T_1)], \tag{30}$$

where T_1 is the time needed to reach the boundary of ${}^{+}R^0$ from the initial point x_1, x_2 with $y = -1$.

In principle, (29) and (25) may be substituted in (30) and from this the derivative ${}^{-}f_{x_2}{}^1(x_1, x_2)$ can be calculated. Analytically the process is extremely laborious for two reasons. First, the number of terms is large. Second, the calculation of the point of intersection of the trajectory with the boundary requires the solution of a transcendental equation. For example, if the intersection is on the line $x_1 = 0$, T_1 must be determined from

$$Ce^{-\rho_3 T_1} + De^{\rho_4 T_1} - r_0 = 0. \tag{31}$$

For these reasons the purely analytic solution of the problem is impractical, and the design must be accomplished with the aid of a digital computer.

Numerical calculations illustrating the procedure were performed for the system $\alpha = 3$, $\beta = 2$. Eqs. (29a–d) were normalized to r_0, and the analytic evaluation of the integral of (30) was obtained. The computer was programmed to calculate first the point at which the trajectory enters the $R_+{}^0$ region. Once this was determined, the integral was calculated from the formula obtained by the analytic integration, ${}^{+}f^0[x_1(T_1), x_2(T_1)]$ was calculated from (25), and finally ${}^{-}f^1[x_1, x_2]$ was found by adding these two (30).

The partial derivative ${}^{-}f_{x_2}{}^1(x_1, x_2)$ which is needed to compute the switching function was obtained as follows. The loss function was computed for a grid of values of

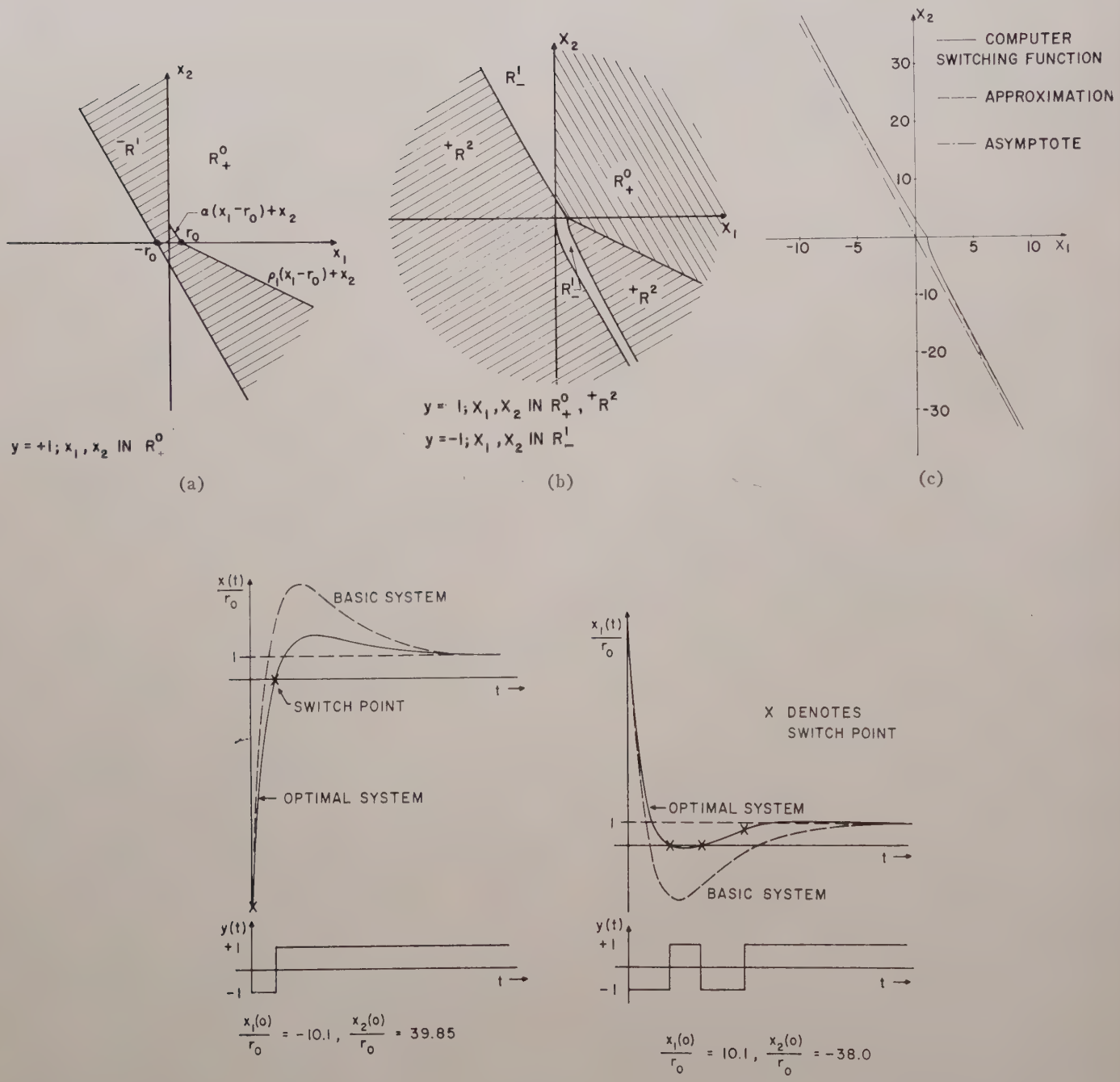


Fig. 3—(a) Location of region R_+° and the potential region R'. (b) Location of regions R_+°, R', ${}^+R^2$. (c) Switching curves for second-order regulator system. (d–e) Typical responses of a second-order regulator system.

x_1, x_2 in ${}^{-}R^1$. The partial derivative was then computed by taking the difference of the loss functions at adjacent points in the phase plane along lines of constant x_1, *i.e.*,

$$ {}^{-}f_{x_2}{}^{1}(x_1, x_2) \cong \frac{1}{\Delta x_2}\left[{}^{-}f^{1}(x_1, x_2 + \Delta x_2) - {}^{-}f^{1}(x_1, x_2)\right]. $$

Since only the sign of this quantity is important, the computer was not programmed to perform the division by Δx_2. In fact, the computer was programmed to search automatically along lines of constant x_1 in ${}^{-}R^1$ until the sign of the partial derivative changed. The value of x_1, x_2 at that point was recorded and the process was then repeated for a new value of x_1. In this way the switching function shown in Fig. 3(c) was obtained.

Using the definition given in (28), the region R^1 is that shown in Fig. 3(b). The shaded areas shown in Fig. 3(b) are the potential ${}^{+}R^2$ regions, and it is necessary next to examine trajectories beginning in this region and passing into R^1.

At this point the analogy with the autonomous case is evident and useful. For very large initial disturbances, *i.e.*, $|x_1|$, $|x_2| \gg 0$, it is clear that the system must behave just as if no input were applied. Thus for $|x_1|$, $|x_2| \gg 0$, the switching function must asymptotically be given by $y = \operatorname{sgn} x_1(\nu x_1 + x_2)$. The latter remark is confirmed by examination of Fig. 3(c). This figure shows that the boundary of the $R_-{}^1$ region approaches a straight line in the phase space for large x_1, x_2. The asymptotic slope of this line is found to be 3.75 ± 0.015.

Thus it is reasonable to expect the line $x_1 = 0$ to remain a switching function and the boundary of the $R_-{}^1$ region shown as the heavy line in Fig. 3(b) to be the other switching function for the regulator system. In other words, the phase plane is divided into the four areas designated R_+, R_- corresponding to the sign of the control variable which is to be used when the system state is in these regions. The fact that this is indeed the optimal switching function was confirmed by programming the computer to calculate the loss function f, using this switching function for the control variable. Accordingly, the sign of $x_1 f_{x_2}(x_1, x_2)$ so calculated corresponded precisely to the sign of the control variable assumed in the region ${}^{+}R^2$, which was the only doubtful region.

The technique described above, of assuming a form for the control function and then checking through (22) as to whether the assumption made of the sign of the control function in a particular region is justified, is a very useful computational procedure.

D. The Response of the Optimal System

The realization of the switching function in a form which is independent of the amplitude of the input signal can be accomplished in many ways. For example, the switching function shown in Fig. 3(b) can be expressed as

$$ y = \operatorname{sgn} u_1 g(u_1, u_2), \tag{32} $$

where

$$ g(u_1, u_2) = \begin{cases} 3.74u_1 + u_2 - \dfrac{3}{1 + 0.329 \log \dfrac{|u_2| + 1}{4}}, & u_1 < 0 \\ 3(u_1 - 1) + u_2, & 0 \le u_1 \le 1 \\ \dfrac{4.3(u_1 - 1)^{0.378}}{1 + 1.15(u_1 - 1)^{0.378}} + 3.74(u_1 - 1) + u_2, & u_1 > 1 \end{cases} \tag{33} $$

and

$$ u_1 = \frac{x_1}{r_0}, \qquad u_2 = \frac{x_2}{r_0}. $$

The expressions for g for $u_1 < 0$, $u_1 > 1$ are approximations for the respective segments of the switching function shown in Fig. 3(b). Fig. 3(c) shows a plot of these expressions in comparison to the computed values. The form of the approximate expressions was chosen to match the asymptotic conditions, the conditions for du_2/du_1 along the switching function $g(u_1, u_2) = 0$, so that there is almost a single change in sign of the function along any straight line in the regions in which each is defined. The particular conditions are

1) $$ \left.\frac{du_2}{du_1}\right|_{u_1=0} = 3, \tag{34} $$

2) $$ \left.\frac{du_2}{du_1}\right|_{u_1=1} = -\infty, \tag{35} $$

3) $$ g(u_1, u_2) \to 3.74u_1 + u_2 \quad u_1, u_2 \to \pm\infty. \tag{36} $$

Condition 1) and the asymptotic condition completely specify ξ in the approximation

$$ g_1(u_1, u_2) = \nu u_1 + u_2 - \frac{\alpha}{1 + \xi \log \dfrac{|u_2| + 1}{\alpha + 1}} \quad u_1 < 0. \tag{37} $$

However the form

$$ g_3(u_1, u_2) = \frac{\nu\xi_1(u_1 - 1)^{\xi_2}}{1 + \xi_1(u_1 - 1)^{\xi_2}} + \nu(u_1 - 1) + u_2 \tag{38} $$

satisfies (35), the asymptotic conditions and the monotonic property for all values of ξ_1, ξ_2 in the ranges $0 < \xi_1$ and $0 < \xi_2 < 1$. The parameters ξ_1, ξ_2 can then be chosen to obtain agreement at two points on the g_3 boundary segment. In particular these were $g_3(2, -5.74) = 0$, which determines ξ_1, and $g_3(8.9, -36.0) = 0$, which then determines ξ_2. The fit of (37) and (38) at the points obtained by the computer are compared in Table I.

TABLE I
TABULATED VALUES OF THE OPTIMAL SWITCHING FUNCTION

x_1	x_2	x_2^*
−10.1	39.91	39.47
− 9.1	36.16	35.74
− 8.1	32.41	32.06
− 7.1	28.79	28.38
− 6.1	25.04	24.66
− 5.1	21.29	21.00
− 4.1	17.54	17.32
− 3.1	13.91	13.69
− 2.1	10.29	10.08
− 1.1	6.66	6.58
− 0.1	3.19	3.30
1.9	− 5.19	− 5.33
2.9	− 9.44	− 9.34
3.9	−13.31	−13.20
4.9	−17.19	−17.05
5.9	−21.06	−20.80
6.9	−24.81	−24.62
7.9	−28.56	−28.43
8.9	−32.31	−32.22
9.9	−35.94	−36.0

* x_2 is the calculated value of the analytic approximation.

Now to complete the design, it is necessary to enable $g(u_1, u_2)$ to apply for all values of r_0. This can be done very simply by writing

$$g(x_1, x_2) = \begin{cases} 3.74x_1 + x_2 - \dfrac{3r_0}{1 + 0.329 \log \dfrac{\left|\dfrac{x_2}{r_0}\right| + 1}{4}}; & x_r \operatorname{sgn} r_0 < 0 \\ 3(x_1 - r_0) + x_2; & 0 \le x_1 \operatorname{sgn} r_0 \le |r_0| \\ 3.74(x_1 - r_0) + x_2 + \dfrac{r_0 |x_1 - r_0|^{0.378}}{|r_0|^{0.378} + 1.15 |x_1 - r_0|^{0.378}}; & x_1 \operatorname{sgn} r_0 > |r_0|. \end{cases} \tag{39}$$

Now the controller for the parameter y must determine its sign continuously according to (32). To do so it is required first to determine the sign of x_1 and then the sign of $g(x_1, x_2)$. To obtain the latter the inequality in (39) which applies at the point x_1, x_2 must first be found and the sign of the corresponding equation must then be calculated.

The response of the optimal system is obtained as a necessary part of the computer solution of the design problem. The method described in Section III-B requires the system trajectory to be computed in all portions of the phase space. For this purpose, the computer was programmed to use the switching function (32), and from this the system trajectories are determined.

Fig. 3(d–e) show the typical response of the adapted system to various initial disturbances. These values were chosen in order to show the maximum number of changes of the parameter value y, and correspondingly the greatest improvement compared to the basic system for initial points in various regions in the phase space.

Inspection of the switching function shows that, regardless of the initial disturbance, the optimal system requires no more than four changes of the parameter value y. This number is obtained, *e.g.*, for the case of $r_0 > 0$, when $x_1 \gg 0$ and x_2 is sufficiently negative. The response of the system under this condition is similar to that shown in Fig. 3(e).

E. Higher-Order Systems with Single Parameter Control

The procedure to determine the switching functions for finite T systems of order greater than two is described below. For nonautonomous systems, it is further assumed that the input can be characterized by a state so that the system equation (1) is applicable.

The switching function for $T^- < \infty$ is time variable as well as state dependent, as mentioned in Section II-B. The rigorous procedure to determine it is carried out as follows. First the trajectories emanating from any point in the phase space ξ corresponding to $y = +K$ and $y = -K$ are computed. Let $\mathbf{x}^+(t; \xi)$ be the trajectory when $y = +k$, and $\mathbf{x}^-(t; \xi)$ be the trajectory when $y = -k$. Corresponding to $\mathbf{x}^+(t, \xi)$, the loss function

$$^+f^0(\xi, T) = \int_0^T \sigma[\mathbf{x}^+(t; \xi)]dt$$

and

$$^-f^0(\xi, T) = \int_0^T \sigma[\mathbf{x}^-(t; \xi)]dt$$

and their derivatives $^+f_\xi{}^0(\xi, T)$, $^-f_\xi{}^0(\xi, T)$ can be computed.

Now let $^+R^0(T)$, $^-R^0(T)$ be defined as

$$^+R^0(T) = \{\xi: {}^+f_\xi{}^0(\xi, T')\boldsymbol{h}(\xi) < 0 \quad \text{for all } T' \le T\}$$

$$^-R^0(T) = \{\xi: {}^-f_\xi{}^0(\xi, T')\boldsymbol{h}(\xi) > 0 \quad \text{for all } T' \le T\}.$$

Then clearly for any T if $\xi \epsilon R_+{}^0(T)$ and if $\mathbf{x}^+(t; \xi) \epsilon R_+{}^0(T-t)$ for all $0 \le t \le T$, then $\operatorname{sgn} y(\xi, T) = +1$, and in fact it remains at $+1$ throughout the remainder of the interval. In this way, the region in the phase space in which $y = +K$ throughout the entire interval T_0 can be found. Call this region $R_+{}^0(T)$. The mathematical definition of the region is

$$R_+{}^0(T) = \{\xi: \xi \epsilon R_+{}^0(T) \text{ and } \mathbf{x}^+(t; \xi) \epsilon R_+{}^0(T - t) \text{ for all } 0 \le t \le T\}.$$

Similarly

$$R_-{}^0(T) = \{\xi: \xi \epsilon R_-{}^0(T) \text{ and } \mathbf{x}^-(t; \xi) \epsilon R_-{}^0(T - t) \text{ for all } 0 \le t \le T\}.$$

Now let $B_+{}^0(T)$ and $B_-{}^0(T)$ be the boundaries of the regions $R_+{}^1(T)$ and $R_-{}^1(T)$, respectively. If the system trajectory begins at a point in the phase space in neither $R_+{}^0(T)$ nor $R_-{}^0(T)$, the parameter value y must switch

at least once in the interval T. To determine those points in phase space requiring just one change in the sign of y, the regions adjacent to $R_+{}^0(T)$ and $R_-{}^0(T)$ are examined.

Suppose the trajectory $\mathbf{x}^-(t;\ \xi)$ where $\xi \epsilon R_+{}^0(T)$, $R_-{}^0(T)$ be such that $\mathbf{x}^-(\tau;\ \xi)\epsilon B_+{}^0(T-\tau)$ for some $\tau < T$. The loss function for this combination of trajectories is

$$^-f^1(\xi,\ T) = \int_0^\tau \sigma[\mathbf{x}^-(t;\ \xi)]dt + {}^+f^0[\mathbf{x}^-(\tau;\ \xi);\ T-\tau]$$

and from this the derivatives $^-f_\xi{}^1(\xi,\ T)$ can be computed. Defining

$$^-R^1(T) = \{\xi:\ ^-f_\xi{}^1(\xi,\ T')\boldsymbol{h}(\xi) > 0 \quad \text{for all } T' \leq T\}$$

the region $R_-{}^1(T)$ in which the sign of y is initially negative and later switches to positive is

$$R_-{}'(T) = \{\xi:\ \xi\epsilon{}^-R^1(T)\ \ \mathbf{x}^-(t;\ \xi)\epsilon{}^-R^1(T-t) \quad \text{for all } 0 \leq t \leq \tau(T)\}.$$

The region $R_+{}^1(T)$ adjacent to $R_-{}^0(T)$, in which sgn $y = +1$ and which requires a single change of the parameter value, can similarly be found. The procedure for determining the regions requiring more than one change of y can similarly be found with the aid of (16). In this way, the entire phase space can be mapped for any interval T.

The procedure described above is extremely laborious even for low-order systems with simple inputs, as has been amply demonstrated previously. A computational procedure which in some ways is simpler than the above procedure is to employ the following recursion method.

Assume any convenient value for the parameter y, for example, $y = +K$. Then, as above, compute the trajectory $x^+(t;\ \xi)$, the loss function $f^0(\xi,\ T)$ for all points in the phase space ξ and for all $T \leq T_0$. The function $f_\xi{}^0(\xi,\ T)h(\xi) = 0$ is then taken as the switching function in the next approximation. Thus in the first approximation, the function

$$y^0(\xi,\ T) = K$$

is used, while in the second approximation

$$y^1(\xi,\ T) = -\ \text{sgn}\, f_\xi{}^0(\xi,\ T)\boldsymbol{h}(\xi)$$

is used. Iterating the procedure, a sequence of approximations

$$y^{n+1}(\xi,\ T) = -\ \text{sgn}\, f_\xi{}^n(\xi,\ T)\boldsymbol{h}(\xi) \qquad n = 0,\ 1,\ \cdots \tag{40}$$

are obtained.

The difference in the two procedures can be described as follows. In the first procedure, the optimal response is built up from the knowledge of the regions in the phase space requiring no switchings in the time T, single switchings and so on until the phase space is completed. In the second method, the optimal switching function is approximated by an assumed function, defined throughout the entire phase space, for all $T \leq T_0$. The optimal switching function is then obtained as a sequence of approximations obtained successively from the loss functions computed for the lower-order approximations. The latter takes no regard of the number of alternations of the sign of y.

The loss functions obtained by the latter process are readily shown to be monotonic decreasing. Since the loss function is obtained by using the trajectory determined by the previous approximation, then

$$f^{n+1}(\mathbf{c},\ T-t) = \int_t^T \sigma[\mathbf{x}(\tau)]d\tau \tag{41}$$

where

$$\dot{\mathbf{x}}(\tau) = \boldsymbol{g}(\mathbf{x}) + y^{n+1}(\mathbf{x},\ T-\tau);$$
$$\mathbf{x}(0) = \mathbf{c} \qquad n = 0,\ 1,\ \cdots.$$

Differentiating (41) with respect to t, it is found that $f^{n+1}(\mathbf{x},\ T)$ satisfies the partial differential equation

$$f_T{}^{n+1} = \sigma(\mathbf{x}) + f_x{}^{n+1}[\boldsymbol{g}(\mathbf{x}) + y^{n+1}\boldsymbol{h}(\mathbf{x})]. \tag{42}$$

Therefore

$$\begin{aligned} f_T{}^{n+1} - f_T{}^n &= (f_x{}^{n+1} - f_x{}^n)\boldsymbol{g}(\mathbf{x}) + y^{n+1}f_x{}^{n+1}\boldsymbol{h}(\mathbf{x}) \\ &\quad - y_n f_x{}^n \boldsymbol{h}(\mathbf{x}) \\ &= (f_x{}^{n+1} - f_x{}^n)\boldsymbol{g}(\mathbf{x}) + y^{n+1}(f_x{}^{n+1} - f_x{}^n)\boldsymbol{h}(\mathbf{x}) \\ &\quad + (y^{n+1} - y^n)f_x{}^n\boldsymbol{h}(\mathbf{x}). \end{aligned} \tag{43}$$

But since y^{n+1} satisfies (40), and hence minimizes $yf_x{}^n\boldsymbol{h}(\mathbf{x})$, the last term on the right hand side of (43)

$$\psi(\mathbf{x},\ T) \triangleq (y^{n+1} - y^n)f_x{}^n\boldsymbol{h}(\mathbf{x})$$

can never be positive, *i.e.*, $\psi(\mathbf{x},\ T) \leq 0$.

Rewriting (43) as

$$(f^{n+1} - f^n)_T = (f^{n+1} - f^n)_x\boldsymbol{g}(\mathbf{x}) + \psi(\mathbf{x},\ T)$$

and comparing this to (42), it is clear that $f^{n+1} - f^n$ satisfies

$$f^{n+1}(\mathbf{C},\ T) - f^n(\mathbf{C},\ T) = \int_0^T \psi[\mathbf{x}(\tau),\ T-\tau]d\tau,$$

where

$$\dot{\mathbf{x}}(\tau) = \boldsymbol{g}(\mathbf{x})$$
$$\mathbf{x}(0) = \mathbf{C}.$$

Since $\psi \leq 0$ throughout the entire phase space, it follows that

$$f^{n+1}(\mathbf{x},\ T) \leq f^n(\mathbf{x},\ T) \qquad n = 0,\ 1,\ \cdots$$

which establishes the monotonic property of the successive approximations.

Appendix I

Necessary and Sufficient Conditions for the Onset of the Switching Mode

The Lagrange method.

Let δx and δy be the variations of the state variables x and the control parameter set y, respectively. These variations are the vectors

$$\delta x \triangleq \begin{bmatrix} \delta x_1 \\ \cdot \\ \cdot \\ \cdot \\ \delta x_n \end{bmatrix}; \qquad \delta y \triangleq \begin{bmatrix} \delta y_1 \\ \cdot \\ \cdot \\ \cdot \\ \delta y_m \end{bmatrix}.$$

Then the variation of the loss function J is given by

$$\delta J = \int_0^T \sigma_x' \delta x dt, \tag{44}$$

where σ_x is the vector derivative of σ given by

$$\sigma_x \triangleq \begin{bmatrix} \dfrac{\partial \sigma}{\partial x_1} \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{\partial \sigma}{\partial x_n} \end{bmatrix}$$

and superscript $'$ indicates transposition.

Notice that because of the constraints of the problem, not all the variations δx and δy are arbitrary. In fact, only the δy variations are arbitrary since having chosen a particular set δy, the variations δx are determined through the system equations (5). A differential equation relating the variations can be obtained from (5) using hypothesis 1 of the theorem, namely that g and ${}_j h$ of the system equation written in the form

$$\dot{x} = g(x) + \sum_{j=1}^{m} {}_j h(x) y_j \tag{45}$$

are C^1 with respect to each component of the state x, and assuming the variations δy are small. The system of equations is

$$\delta \dot{x} = A \delta x + \sum_{j=1}^{m} {}_j h \delta y_j, \tag{46}$$

where A is the $n \times n$ matrix given by

$$A = g_x + \sum_{j=1}^{m} {}_j h_x y_j$$

and

$$g_0 \triangleq \begin{bmatrix} \dfrac{\partial g_1}{\partial x_1} & \cdots & \dfrac{\partial g_1}{\partial x_n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \dfrac{\partial g_n}{\partial x_1} & \cdots & \dfrac{\partial g_n}{\partial x_n} \end{bmatrix} = \left[\frac{\partial g_i}{\partial x_j} \right]$$

and similarly

$${}_k h_x = \left[\frac{\partial_k h_i}{\partial x_j} \right] \qquad k = 1, \cdots, m.$$

In (46) higher-order variations have been neglected so that A contains only known time functions; that is, in taking variations, the optimizing solutions for the state x and the parameter set y are assumed known and these solutions are perturbed by the variations δx and δy. Under these assumptions, neglecting higher-order variations amounts to assuming a complete knowledge of A on the interval $0 \leq t \leq T$. Similarly, ${}_j h_x$ would also be a set of known time functions on this interval.

Under these assumptions the differential equation (47) is a linear set whose solution can be described explicitly by well-known conventional methods.[9] In fact if $Y(t)$ is the solution of the matrix differential equation

$$\dot{Y}(t) = A(t) Y(t) \tag{47}$$

with the initial condition $Y(0) = I$, where I is the identity matrix, then the solution of (47) can be written as

$$\delta x = Y(t) c + \sum_{j=1}^{m} \int_0^t Y(t) Y^{-1}(\tau) {}_j h(\tau) \delta y_j(\tau) d\tau, \tag{48}$$

where c is the initial condition of δx. Since all admissible solutions meet the boundary condition $x(0) = x_0$, it follows that $\delta x(0) = 0$, hence $c = 0$. Thus the solution of (47) meeting the proper boundary condition is

$$\delta x = \sum_{j=1}^{m} \int_0^t Y(t) Y^{-1}(\tau) {}_j h(\tau) \delta y_j(\tau) d\tau. \tag{49}$$

The variation δJ can now be written entirely in terms of the independent arbitrary variations δy by substituting (49) in (44). This yields

$$\delta J = \sum_{j=0}^{m} \int_0^T \int_0^t \sigma_x'(t) Y(t) Y^{-1}(\tau) {}_j h(\tau) \delta y_j(\tau) d\tau dt.$$

Finally, interchanging the order of integration and also the variables t and τ enables the equation to be written in terms of the solution on the latter part of the interval and further places the variations δy_j as a multiplying factor in the last indicated integration, as follows:

$$\delta J = \sum_{j=1}^{m} \int_0^T \int_t^T \sigma_x'(\tau) Y(\tau) Y^{-1}(t) {}_j h(t) \delta y_j(t) d\tau dt.$$

Now consider that at some time t, the optimal value of the parameter y_j is k_j. In order that the perturbed system $y_{j\,\mathrm{opt}} + \delta y_j = k_j + \delta y_j$ satisfy the boundedness constraint, the variation δy_j can only be negative. (δy_j is always sufficiently small so that the lower bound is also not exceeded.) Similarly if $y_{j\,\mathrm{opt}} = -k_j$, then only positive variations for δy_j are admissible. If $y_{j\,\mathrm{opt}}$ is within the bound, the variation can be arbitrary.

[9] Coddington and Levinson, *op. cit.*, chap. 2.

Since perturbing an optimal solution can never decrease δJ, it follows that $\delta J \geq 0$ for all admissible variations. In particular, when y_j is at the bound, the inequality may hold. However, when y_j is within the bound, the equality must hold. Taking these permissible variations into account then, it is *necessary* that

$$\int_t^T \sigma_x'(\tau)\, Y(\tau)\, Y^{-1}(t)\, {}_j\boldsymbol{h}(t)\, d\tau \begin{cases} \leq 0; & y_j(t) = k_j \quad (50a) \\ = 0; & |\, y_j(t)\,| < k_j \quad (50b) \\ \geq 0; & y_j(t) = -k_j \quad (50c) \end{cases}$$

under the conditions listed.

The next step in the proof consists of obtaining equations describing the types of systems, the types of criteria σ and the conditions under which these systems and criteria make the equality relation above hold. Since $Y(\tau)$ requires complete knowledge of $\boldsymbol{y}$ on the interval $0 \leq \tau \leq T$, (50) are not useful in themselves in determining the switching conditions. Because of the inequalities, the necessary conditions which define the times at which switching occurs cannot be expressed directly in terms of the known functions A, ${}_j\boldsymbol{h}$ and σ. Nonetheless, (50b) can be used to determine the conditions for the onset of the switching mode as indicated next.

Consider that $|\, y_j(t)\,| < k_j$ on some time interval. Then not only is (50b) valid, but it also possesses a derivative with respect to t. Let

$$\boldsymbol{z}'(t) \triangleq \int_t^T \sigma_x'(\tau)\, Y(\tau)\, Y^{-1}(t)\, d\tau. \qquad (51)$$

Then

$$\dot{\boldsymbol{z}}'(t) = -\sigma_x'(t) + \int_t^T \sigma_x'(\tau)\, Y(\tau)\, Y^{-1}(t)\, d\tau.$$

Since[10]

$$Y^{-1}(t) = -Y^{-1} Y Y^{-1},$$

substituting (47) in this equation results in

$$Y^{-1}(t) = -Y^{-1} A.$$

Hence

$$\dot{\boldsymbol{z}}'(t) = -\sigma_0'(t) - \int_t^T \sigma_x'(\tau)\, Y(\tau)\, Y^{-1}(t)\, A(t)\, d\tau.$$

From the definition (51), it follows that

$$\dot{\boldsymbol{z}}' = -\sigma_x' - \boldsymbol{z}' A. \qquad (52)$$

Thus $\boldsymbol{z}$ is the solution of the set of differential equations

$$\dot{\boldsymbol{z}} = -A'\boldsymbol{z} - \sigma_0$$

with the boundary condition $\boldsymbol{z}(T) - 0$.

In terms of the vector $\boldsymbol{z}$, (50b) can be expressed as

$$\boldsymbol{z}'\, {}_j\boldsymbol{h} = 0. \qquad (53)$$

Differentiating the above equation with respect to t yields

$$\dot{\boldsymbol{z}}\, {}_j\boldsymbol{h} + \boldsymbol{z}'\, {}_j\dot{\boldsymbol{h}} = 0$$

and substituting (52) in this equation results in

$$\sigma_x'\, {}_j\boldsymbol{h} = \boldsymbol{z}'({}_j\dot{\boldsymbol{h}} - A\, {}_j\boldsymbol{h}).$$

Now

$$\sigma_x'\, {}_j\boldsymbol{h} = 0, \qquad j = 1, 2, \cdots, m \qquad (54)$$

if the parameter control has no immediate effect on the state components measured by σ. This means that if x_i is an argument of σ, then $\dot{x}_i$ is independent of y_j. If σ is a measure of the first p derivatives of the system error, the condition is satisfied if the $p+1$ error derivative is independent of y_j. The condition (54) is then exactly the same as that stated in hypothesis 2, which was used previously in the dynamic programming derivation to establish the need for a switching mode.

It follows then that for the systems considered, the necessary condition in a region in which $|\, y_j\,| < k_j$ is that

$$\boldsymbol{z}'({}_j\dot{\boldsymbol{h}} - A\, {}_j\boldsymbol{h}) = 0. \qquad (55)$$

In other words, an optimal system produces a parameter value y_j smaller in magnitude than k_j at those values of the state $\mathbf{x}$ satisfying both (53) and (54).

Now (55) is satisfied if either

$$\boldsymbol{z} = 0 \qquad (56)$$

of if

$${}_j\dot{\boldsymbol{h}} = A\, {}_m\boldsymbol{h} + \boldsymbol{m}, \qquad (57)$$

where $\boldsymbol{m}$ is any vector orthogonal to $\boldsymbol{z}$.

Consider now the condition (56). Evidently if (56) is satisfied, so is (53). Since $Y(t)$ and $Y^{-1}(t)$ are never singular matrices,[11] using the definition (51) in (56) gives the necessary condition as

$$\int_t^T \sigma_{\mathbf{x}}'\, Y(\tau)\, d\tau = 0. \qquad (58)$$

It is important to note that (58) applies for any $j = 1, \cdots, m$ of the control variables $\boldsymbol{y}$. It will now be shown that the condition (57) yields exactly the same relation, so that (58) serves to define the region in which the parameter values need not lie on their respective bounds.

Eq. (57) is exactly the same as (46) with ${}_j\boldsymbol{h}$ replacing $\mathbf{x}$ and $\boldsymbol{m}$ replacing

$$\sum_{j=1}^{m} {}_j\boldsymbol{h} y_j.$$

Hence, from (47) it is clear that

$${}_j\boldsymbol{h} = Y(T)\mathbf{c} + \int_0^t Y(t)\, Y^{-1}(\tau)\boldsymbol{m}(\tau)\, d\tau, \qquad (59)$$

[10] Coddington and Levinson, *op. cit.*, ch. 2, p. 70.

[11] Coddington and Levinson, *op. cit.*, ch. 2, pp. 69–71.

where $\mathbf{c}$ is a constant vector. Since any solution of (57) must satisfy (55), the vector $\mathbf{c}$ is arbitrary. Thus substituting (59) and (51) in (53) leads to

$$\int_t^T \sigma_{-x}' Y(\tau) c d\tau + \int_t^T \int_0^t \sigma_x'(\tau_1) Y(\tau_1) Y^{-1}(\tau_2) \boldsymbol{m}(\tau_2) d\tau_1 d\tau_2 = 0. \quad (60)$$

But (60) must be true for an arbitrary vector $\mathbf{c}$. Hence (60) is satisfied if and only if each member of the left hand side of (60) is itself zero. Thus *condition* (57) *is exactly the same as* (58). Note that no new conditions are obtained from the second member of (60) since one of its factors yields (58).

The existence of a region in which the parameter value is continuously varied is now defined by those regions in which (58) is satisfied. Consider now that (58) is met at time t. Then (58) can be written as

$$\int_t^T \sigma_x' Y(\tau) d\tau = \int_t^{t+\Delta} \sigma_x' Y d\tau + \int_{t+\Delta}^T \sigma_x' Y d\tau = 0. \quad (61)$$

For sufficiently small Δ, the system remains in the continuous mode region. It follows that

$$\int_{t+\Delta}^T \sigma_x' Y d\tau = 0 \quad (62)$$

whereas

$$\int_t^{t+\Delta} \sigma_x' Y d\tau \xrightarrow[\Delta \to 0]{} \sigma_x' Y(t) \Delta. \quad (63)$$

Since $Y(t)$ is never singular, substituting (62) and (63) in (61) shows that *the necessary condition is*

$$\sigma_x = 0 \quad (64)$$

on the interval. Since the vector σ_x must be the null vector, *the derivatives of the measurement function with respect to each state component must be separately set to zero.*

In particular, suppose the system error can be expressed in terms of $\mathbf{x}$ as $\epsilon = \boldsymbol{a}'\mathbf{x}$ where $\boldsymbol{a}$ is a constant vector. Then if the loss function is a measure of the error alone, *i.e.*, has the form (2), namely

$$J = \int_0^T G(\epsilon) dt,$$

then

$$\sigma(\mathbf{x}) = G(\boldsymbol{a}'\mathbf{x}).$$

It follows that

$$\sigma_x = \frac{dG(\epsilon)}{d\epsilon} \boldsymbol{a}.$$

The requirement (54) that

$$\sigma_x' {}_j\boldsymbol{h} \equiv 0, \qquad j = 1, \cdots, m \quad (54)$$

for all $\mathbf{x}$ simply means that $\boldsymbol{a}'{}_j\boldsymbol{h} = 0$, $j = 1, \cdots, m$. In other words, the control functions $\boldsymbol{y}$ exert only differential control on the system output. Note that

$$\dot{\epsilon} = \boldsymbol{a}'\dot{\mathbf{x}} = \boldsymbol{a}'\left(\boldsymbol{g} + \sum_{j=1}^m {}_j\boldsymbol{h} y_j\right) = \boldsymbol{a}'\boldsymbol{g}$$

so that the first derivative of the error is continuous, even where the y_j's are discontinuous.

The necessary (and sufficient) condition, namely $\sigma_x = 0$, means in this case that $dG(\epsilon)/d\epsilon = 0$. If $G(\epsilon)$ is a positive definite monotonic increasing function of the magnitude of ϵ of class C^1, it follows that $\epsilon \equiv 0$ everywhere in the interval in which (64) applies. Hence, during this interval, the output reproduces the system input identically.

In this section it has been shown that periods in which the control parameters $\boldsymbol{y}$ may have values within their respective limits may exist. But once such a period is entered, it is terminated only when the input can no longer be reproduced identically with the control parameter within their prescribed limits. This means that, even though it was known *a priori* that the input could be reproduced identically only on the interval $0 \leq \tau = t_1 < T$, and to do so on the interval $t_1 < \tau \leq T$ would require the magnitude of at least one among the control parameters $\boldsymbol{y}$ to exceed its bound, an optimal system would not begin its switching mode of operation prior to time t_1.

Acknowledgment

The author wishes to thank R. Schwarz for his counsel and encouragement during the course of this work.

A Minimal Time Discrete System*

C. A. DESOER†, SENIOR MEMBER, IRE, AND J. WING†, MEMBER, IRE

Summary—Consider a sampled-data control system with the following sequence of components in the forward path: a sampler with period T, a zero-order hold circuit, a linear amplifier with saturation limits ± 1, and a plant with transfer function

$$G(s) = \frac{1}{\prod_{i=1}^{n}(s-\lambda_i)}.$$

It is assumed that the poles $\lambda_1, \lambda_2, \cdots, \lambda_n$ of $G(s)$ are real, distinct, and non-positive (a single integral is permissible). The sampler, zero-order hold, and saturating amplifier constrain $f(t)$, the forcing function of $G(s)$, to be piecewise constant with values between -1 and $+1$. The forcing function $f(t)$ is completely defined, for $t>0$, by the sequence of numbers $f_1, f_2, \cdots$, where f_i is the value of $f(t)$ during the i'th sampling period.

The minimal time regulator problem for the above system can then be stated as follows: Given $G(s)$ with an arbitrary set of initial conditions [*i.e.*, the state vector $\vec{c}(0)$ defined by its components $c(0), \dot{c}(0), \cdots, c^{n-1}(0)$]; find the forcing function $f(t)$ [specified by $f_1, f_2, \cdots$ and satisfying $|f_i| \leq 1$], and the corresponding computer in the feedback loop which will bring the system to equilibrium in the minimum number of sampling periods. Any such forcing function will be called an optimal control.

The first step is to consider R_N' the set of all initial states $\vec{c}(0)$ from which the origin can be reached in N sampling periods or less. From this definition all such states are characterized algebraically and geometrically: R_N' is shown to be a convex polyhedron with

$$2\sum_{k=1}^{n}\binom{N-1}{k-1}$$

vertices.

Let R_N be the set of all initial states $\vec{c}(0)$ from which the origin can be reached in N sampling periods and no less. Each point of R_N is shown to have a unique canonical representation. The coefficients appearing in the canonical representation suggest an optimal control.

To obtain this particular optimal control we define a surface in state space called the critical surface. It is shown that this optimal control will be generated by the following procedure: at the beginning of each sampling period the distance ϕ from the state of the system to the critical surface is measured along a fixed specified direction; if $\phi \geq 1$ (or ≤ -1) then the forcing function for that sampling period is $+1$ (or -1); if $|\phi| < 1$, then the forcing function is ϕ. For a third-order plant it is shown that the critical surface has certain properties which lead to a simple analog computer simulation.

I. Introduction

In recent years a considerable amount of work has been given to the problem of optimal control systems, especially for the continuous case [1]–[9]. In these problems, the key aspect is that the control signals are restricted to belong to a closed and bounded set; in the case of a single control signal u it is restricted by the condition $|u(t)| \leq \gamma$ where γ is a prescribed number that plays a pre-eminent role in the design. The case of discrete systems has received less attention [10], [11].

This paper considers the following sampled-data control system: The forward path consists of a sampler with period T, a zero-order hold circuit, a linear amplifier and a plant with transfer function

$$G(s) = \frac{1}{\prod_{i=1}^{n}(s-\lambda_i)}.$$

The poles of $G(s)$ must be real, distinct, and nonpositive. Thus, a single integrator is permissible. The control signal $f(t)$ applied to the sampler is restricted by $|f(t)| \leq 1$. The problem is: Given an arbitrary set of initial conditions, find the forcing function $f(t)$, satisfying $|f(t)| \leq 1$, and the corresponding computer to be placed in the feedback loop which will bring the system to equilibrium in the minimum number of sampling periods.

The principal purpose of this paper is to establish carefully the optimality of the proposed control function and to indicate how a computer can be built to generate this optimal control function.

In Section II the problem is carefully stated and then formulated in new variables so that the analysis in the remainder of the paper is as simple as possible. The principal result of this section is (8), which is the state transition equation of the system. In Section III, the set R_N' is defined as the set of all states from which the origin can be reached in N sampling periods or less. In equation form, R_N' is given by (12). From (12) a set of vertex points V_N is defined. It is then shown that R_N' is the convex hull of V_N; hence R_N' is a convex polyhedron. Finally a procedure is obtained for writing down all the points of V_N (Corollary 2). Next the set R_N is defined as the set of all states from which the origin can be reached in N sampling periods and no less. A unique canonical representation is then given for all points of R_N [see (17) and (18)]. In Section IV the critical surface is defined. This surface plays, in the generation of the optimal control, a role which is similar but not identical to that of the switching surface in the continuous case. Finally, in Section V the method for generating an optimal control is established. It is as follows:

Let γ be the state of the system at the beginning of the present sampling period. Let ϕ be the distance between γ and the critical surface measured in the direction of r_1 and using the length of r_1 as the unit length. The optimal control is then sat(ϕ).[1] Section VI shows, by using two properties of the critical surface, that a

* Received by the PGAC, December 15, 1960.

† Elec. Engrg. Dept., University of California, Berkeley. This research was jointly supported by the Electronics Res. Directorate of the AF Cambridge Res. Center, Air. Res. and Dev. Command, under Contract AF 19(604)-5460 and the Air Force Office of Scientific Res. of the Air Res. and Dev. Command, under Contract No. AF 18(1600)-1521.

[1] sat $(x) \equiv x$ if $|x| < 1$ and $\equiv x/|x|$ when $|x| \geq 1$.

conceptually simple analog computer using standard analog computer techniques can be constructed to generate the proposed optimal control function. Finally, a numerical example is given to illustrate the fact that the canonical representation of R_N is just the sequence generated by the analog computer.

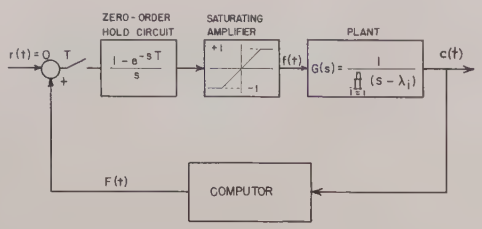


Fig. 1—Block diagram of servomechanism.

II. Statement of the Problem

Fig. 1 shows the servomechanism that will be considered throughout the paper. The plant is characterized by the stable transfer function $G(s)$. We assume that the poles $\lambda_1, \lambda_2, \cdots, \lambda_n$ of $G(s)$ are *real*, *distinct* and *nonpositive*, *i.e.*, one of them may be zero. The actual forcing function $f(t)$, which is the output of the saturating amplifier, must at all times satisfy

$$|f(t)| \leq 1. \tag{1}$$

The sampler and the zero-order hold circuit require that $f(t)$ be piecewise constant: $f(t)=f_k$ for $(k-1)T<t\leq kT$, $(k=1, 2, \cdots)$. Therefore, (1) becomes

$$|f_k| \leq 1, \qquad (k = 1, 2, \cdots). \tag{2}$$

The forcing function $f(t)$ is completely defined, for $t>0$, by the sequence of numbers $f_1, f_2, \cdots$.

The problem which we propose to solve is the following: Assuming that the input $r(t)$ is zero at all times and given an arbitrary set of initial conditions $[c(0), \dot{c}(0), \cdots, c^{(n-1)}(0)]$ find the forcing function $f(t)$, [specified by $f_1, f_2, \cdots$ and satisfying (2)] and the corresponding computer which will bring the system to equilibrium in the minimum number of sampling periods.

As an initial step in the solution of this problem, the state transition matrix equation for a sampling period of T will now be derived. Let $\mathbf{c}(t)$ be the column vector whose components are $c(t)$, $c^{(1)}(t), \cdots, c^{(n-1)}(t)$. Let the denominator of $G(s)$ be $s^n-a_{n-1}s^{n-1}-\cdots-a_1s-a_0$. The differential equation of the system can be written in the standard matrix form

$$\dot{\mathbf{c}}(t) = \boldsymbol{B}\mathbf{c}(t) + f_k\boldsymbol{a} \quad \text{for} \quad (k-1)T < t \leq kT \qquad k = 1, 2, \cdots, \tag{3}$$

where

$$\boldsymbol{B} = \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot \\ 0 & 0 & 1 & \cdot & \cdot \\ & & & 0 & 1 \\ a_0 & a_1 & \cdot & \cdot & a_{n-1} \end{bmatrix} \quad \boldsymbol{a} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

The eigenvalues of $\boldsymbol{B}$ are the roots of the denominator of $G(s)$ and will be assumed to be ordered as follows:

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n \leq 0. \tag{4}$$

The solution of (3) for a sampled system of period T is

$$\mathbf{c}_k = e^{\boldsymbol{B}T}\mathbf{c}_{k-1} + f_k\boldsymbol{b} \qquad (k = 1, 2, \cdots), \tag{5}$$

where

$\mathbf{c}_k$ = state of the system at time $t=kT$,
f_k = forcing function during the kth sampling period, $(k-1)T<t<kT$,
$\boldsymbol{b}=(I-e^{\boldsymbol{B}T})\boldsymbol{B}^{-1}\boldsymbol{a}$.[2]

Observe that 1) the eigenvalues of $e^{\boldsymbol{B}T}$ are

$$e^{\lambda_i T} \quad (i = 1, 2, \cdots, n)$$

with eigenvector $\boldsymbol{u}_i=\text{col}(1, \lambda_i, \lambda_i^2, \cdots, \lambda_i^{n-1})$, and 2) these eigenvectors constitute a basis for the state space, hence $\det(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n)\neq 0$. To simplify the algebra required by the use of (5), use the eigenvectors $\boldsymbol{u}_i$ as a basis, thus

$$\mathbf{c}_k = \sum_{i=1}^{n} \gamma_{ki}\boldsymbol{u}_i$$

or

$$\mathbf{c}_k = \boldsymbol{U}\boldsymbol{\gamma}_k, \text{ where } \boldsymbol{\gamma}_k = \text{col}\,(\gamma_{k1}, \gamma_{k2}, \cdots, \gamma_{kn}), \tag{6}$$

where the transformation matrix $\boldsymbol{U}$ has $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n$ as columns. Similarly,

$$\boldsymbol{b} = \sum_{i=1}^{n} d_i\boldsymbol{u}_i$$

or

$$\boldsymbol{b} = \boldsymbol{U}^{-1}\boldsymbol{d} \quad \boldsymbol{d} = \text{col}\,(d_1, d_2, \cdots, d_n). \tag{7}$$

With the transformation of variables given by (6), (5) becomes

$$\boldsymbol{\gamma}_k = \boldsymbol{\Lambda}\boldsymbol{\gamma}_{k-1} + f_k\boldsymbol{d}, \tag{8}$$

where [13]

$\boldsymbol{\Lambda}$ = diag $(\exp \lambda_1 T, \exp \lambda_2 T, \cdots, \exp \lambda_n T)$,
$\boldsymbol{\gamma}_k$ = state of the system at time $t=kT$, *i.e.*, at the end of kth sampling period $(k-1)T<t\leq kT$,
f_k = forcing function during the kth sampling period.

Eq. (8) is the desired transition matrix equation which will be utilized in the remainder of the paper: it gives the state of the system at the end of the kth sampling period ($\boldsymbol{\gamma}_k$) in terms of the state at the beginning of the kth sampling period ($\boldsymbol{\gamma}_{k-1}$) and the forcing function (f_k) for the duration of the kth sampling period.

[2] It must be stressed that the matrix $e^{\boldsymbol{B}T}$ is equal to a polynomial in $\boldsymbol{B}$ of degree $(n-1)$. (See [13], p. 121.)

For the system under consideration it is easily verified that the conditions of controllability [12] are satisfied: the n vectors $\boldsymbol{B}^k\boldsymbol{a}$, $(k=0, 1, 2, \cdots, n-1)$ are linearly independent. These conditions are equivalent to requiring $\boldsymbol{\Lambda}^k\boldsymbol{d}$ $(k=0, 1, 2, \cdots, n-1)$ to be linearly independent, *i.e.*, $d_i \neq 0$ $(i=1, 2, \cdots, n)$.

III. Characterization of R_N'

A. Definition of R_N'

R_N' is defined as the set of all states from which the origin can be reached in N sampling periods or less. That is, for every point $\boldsymbol{\gamma}_0$ in R_N' there is a sequence of forcing functions, $f_1, f_2, \cdots$, satisfying (2), which will make $\boldsymbol{\gamma}_N = \boldsymbol{0}$. With these forcing function and using (8),

$$\boldsymbol{\gamma}_0 = -f_1\boldsymbol{\Lambda}^{-1}\boldsymbol{d} - f_2\boldsymbol{\Lambda}^{-2}\boldsymbol{d} - \cdots - f_N\boldsymbol{\Lambda}^{-N}\boldsymbol{d} \tag{9}$$

with $|f_i| \leq 1$, $i=1, 2, \cdots, N$. For convenience, the following notation is introduced:

$$\boldsymbol{r}_k = -\boldsymbol{\Lambda}^{-k}\boldsymbol{d} \qquad (k = 1, 2, 3, \cdots), \tag{10}$$

and (9) becomes

$$\boldsymbol{\gamma}_0 = \sum_{i=1}^{N} f_i\boldsymbol{r}_i \quad \text{with} \quad |f_i| \leq 1 \ (i = 1, 2, \cdots, N). \tag{11}$$

Thus one can state compactly an equivalent definition of R_N' as follows:

$$R_N' = \left\{\boldsymbol{\gamma} \,\middle|\, \boldsymbol{\gamma} = \sum_{i=1}^{N} f_i\boldsymbol{r}_i, \ |f_i| \leq 1, \ i = 1, 2, \cdots, N\right\}. \tag{12}$$

B. Properties of the Vectors $\boldsymbol{r}_k$

Before proceeding to the complete characterization of R_N', let us indicate three properties of the vectors $\boldsymbol{r}_k$.

Property 1: For any set of distinct indexes $i_1, i_2, \cdots, i_n$, the vectors $\boldsymbol{r}_{i_1}, \boldsymbol{r}_{i_2}, \cdots, \boldsymbol{r}_{i_n}$ are linearly independent.

Property 2: Consider n arbitrary indexes such that

$$\alpha_1 < \alpha_2 < \cdots < \alpha_n < N + 1.$$

Then

a) the ξ_{α_k} satisfying the equality

$$\boldsymbol{r}_{N+1} = \sum_{k=1}^{n} \xi_{\alpha_k}\boldsymbol{r}_{\alpha_k}$$

are uniquely determined;

b) $\operatorname{sgn} \xi_{\alpha_n} = +1$;

c) $\operatorname{sgn} \xi_{\alpha_{k+1}} = -\operatorname{sgn} \xi_{\alpha_k}$ $(k=1, 2, \cdots, n-1)$.

Another way of expressing facts b) and c) is to say that the sequence $\{\xi_{\alpha_k}\}$ is of alternating sign and ξ_{α_n} is positive, or equivalently that the sequence $\{\xi_{\alpha_k}\}$ of n elements has $n-1$ sign variations.

Property 3: If, for some arbitrary integer k,

$$\sum_{i=1}^{k} \xi_i\boldsymbol{r}_{v_i} = 0,$$

where $v_1 < v_2 < \cdots < v_k$, then the sequence $\{\xi_i\}$ has at least n sign variations.

These three properties are proved in Appendix I. They constitute the key to the characterization of R_N' and, consequently, to the determination of the optimal control.

C. Definition of the Convex Hull of V_N, $C(V_N)$

Going back to the characterization of R_N', let us observe the obvious:

$$R_1' = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma} = f_1\boldsymbol{r}_1, \ |f_1| \leq 1\},$$

$$R_2' = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma} = f_1\boldsymbol{r}_1 + f_2\boldsymbol{r}_2, \ |f_i| \leq 1, i = 1, 2\},$$

$$R_n' = \{\boldsymbol{\gamma} \mid \boldsymbol{\gamma} = f_1\boldsymbol{r}_1 - f_2\boldsymbol{r}_2 + \cdots + f_n\boldsymbol{r}_n, \ |f_i| \leq 1, \ i = 1, 2, \cdots, n\}. \tag{13}$$

R_1' is a line segment parallel to $2\boldsymbol{r}_1$, centered at the origin. R_2' is a parallelogram centered on the origin and whose edges are parallel to $2\boldsymbol{r}_1$, $2\boldsymbol{r}_2$. R_n' is an n-dimensional parallelepiped centered on the origin, whose edges are parallel to $2\boldsymbol{r}_1, 2\boldsymbol{r}_2, \cdots, 2\boldsymbol{r}_n$. For the case considered in the numerical example (Section VII), the regions R_1' to R_4' are illustrated by Figs. 11 to 13.

To characterize R_N', when $N>n$, is slightly more complicated and requires a little background. For this purpose, the set V_N is introduced and is defined, for all $N>n$, as follows:

V_N is the set of points in n space of the form

$$\sum_{i=1}^{N} \epsilon_i\boldsymbol{r}_i, \tag{14}$$

where: 1) each ϵ_i is equal to 1 in absolute value; 2) the sequence $\{\epsilon_i\} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_N)$ consists of at most n subsequences of consecutive ϵ_i's that have the same sign and each subsequence has a sign opposite to that of the preceding one and of the following one. (Equivalently, the ϵ-sequence cannot have more than $n-1$ sign variations.)

For the cases $n=3$ and $N=4$, the ϵ-sequences are illustrated in Fig. 2: when $\epsilon_i = \pm 1$, a square block of height ± 1 is drawn. It is easy to check that in n space, V_N has

$$2\sum_{k=1}^{n}\binom{N-1}{k-1}$$

points; obviously, this number increases much faster than N. Observe also that when $N=n$, V_N consists of 2^n points: the ϵ-sequences are the 2^n possible n-tuples of $+1$ and -1.

A new idea must now be introduced: the concept of convex hull. By definition, $C(V_N)$, *the convex hull of* V_N [14], is the smallest convex set containing V_N; or equivalently, if $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_r$ (r arbitrary) are the points of V_N, then

$$\mathbf{x} = \sum_{i=1}^{r} \mu_i\mathbf{x}_i$$

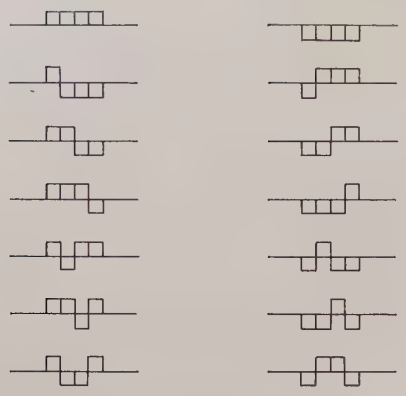


Fig. 2—Complete listing of the ϵ sequences of V_4 for $n=3$.

is a point of $C(V_N)$, provided

1) $\mu_i \geq 0$, 2) $\sum_{i=1}^{r} \mu_i = 1$.

Physically, the convex hull of V_N, $C(V_N)$, is the set of all points $\mathbf{x}$ which are the centers of gravity of the points $\mathbf{x}_i$ of V_N when $\mathbf{x}_i$ is assigned any mass μ_i obeying 1) and 2).

A basic property of convex hulls that will be used is the following [14].

Every point $\boldsymbol{\gamma}$ of $C(V_N)$ is representable in the form

$$\boldsymbol{\gamma} = \sum_{i=0}^{n} \mu_i \mathbf{y}_i, \ \mu_i \geq 0, \qquad \sum_{i=0}^{n} \mu_i = 1 \tag{15}$$

where the $\mathbf{y}_i$'s belong to V_N but the selection of the $\mathbf{y}_i$'s depends on $\boldsymbol{\gamma}$.

If $\boldsymbol{\gamma}$ is a boundary point one may take $\mu_0 = 0$.[3] For example, in 3-space, a cube is the convex hull of the set of its vertices, hence any point in the cube is the center of gravity of 4 appropriately chosen vertices; any point on the boundary is the center of gravity of 3 appropriately chosen vertices. Let us introduce, parenthetically, a fact which will help the visualization of later results. First, a definition [14]: a point $\mathbf{x}$ of $C(V_N)$ is a *vertex point* (or extreme point) if there are no points $\mathbf{x}_1$ and $\mathbf{x}_2$ of $C(V_N)$ with $\mathbf{x}_1 \neq \mathbf{x}_2$, with the property that $\mathbf{x} = \mu\mathbf{x}_1 + (1-\mu)\mathbf{x}_2$ for some μ, $0 < \mu < 1$.

From this definition and the definition of $C(V_N)$ and V_N it is easily verified that:

$$\text{all points of } V_n \text{ are vertices of } C(V_n). \tag{16}$$

It should be stressed that this property is not true for larger values of N, and that this is the main reason why the characterization of the vertices of R_N' is not easy.

D. Fundamental Theorem and Classification of the Points of V_N

The characterization of R_N' is completed by (13) and the following theorem:

Theorem 1: Let V_N be the set of points defined by (14), and let $C(V_N)$ be the convex hull of the set V_N; then $R_N' = C(V_N)$ for all $N \geq n$.

The proof is by induction and is given in detail in Appendix II. Note that from what has already been said (13), $R_n' = C(V_n)$, *i.e.*, the theorem is true for $N = n$. The induction assumption is $R_N' = C(V_N)$, from which one proves that $R_{N+1}' = C(V_{N+1})$. The proof is based on an idea which consists in classifying the points of V_N.

Let δ be any positive number small enough so that the η_i's defined by the expansion

$$\mathbf{r}_{N+1} = \sum_{i=1}^{n} \eta_i \mathbf{r}_{a_i} \qquad a_1 < a_2 < \cdots < a_n < N+1$$

satisfy the inequality $|\delta\eta_i| < 1$ for all i and for all possible choices of $a_1, a_2, \cdots, a_n$. The classification of the points of V_N is as follows:

V_N^+ contains all the points $\mathbf{p}_i$ of V_N such that $\mathbf{p}_i + \delta\mathbf{r}_{N+1} \notin C(V_N)$ but $\mathbf{p}_i - \delta\mathbf{r}_{N+1} \in C(V_N)$;
V_N^- contains all the points $\mathbf{q}_i$ of V_N such that $\mathbf{q}_i - \delta\mathbf{r}_{N+1} \notin C(V_N)$ but $\mathbf{q}_i + \delta\mathbf{r}_{N+1} \in C(V_N)$;
V_N^b contains all the points $\mathbf{b}_i$ of V_N such that $\mathbf{b}_i = \delta\mathbf{r}_{N+1} \notin C(V_N)$ and $\mathbf{b}_i + \delta\mathbf{r}_{N+1} \notin C(V_N)$.

This classification is illustrated in Fig. 3 where $C(V_N)$ is shown as a parallelogram. Roughly speaking, with respect to the direction $\mathbf{r}_{N+1}$, V_N^+ consists of all the points which are on the top of $C(V_N)$, V_N^- those on the bottom and V_N^b those on the edges (sides).

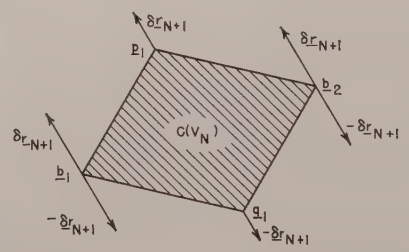


Fig. 3—Classification of the points of V_N.

Each of the above classes is non-empty and can be easily described: 1) V_N^+ contains all the points of V_N whose ϵ sequence has $n-1$ changes of sign and whose last subsequence is of positive sign, 2) V_N^- contains all the points of V_N whose ϵ sequence has $n-1$ changes of sign and whose last subsequence is of negative sign, 3) V_N^b contains all the points of V_N whose ϵ sequence has $n-2$ changes of sign or less.

Let us justify statement 1. It is asserted that if $\mathbf{p}$ satisfies the conditions stated in 1, then $\mathbf{p} + \delta\mathbf{r}_{N+1} \notin C(V_N)$ and $\mathbf{p} - \delta\mathbf{r}_{N+1} \in C(V_N)$.

Recall property 2, take each a_i in a different ϵ subsequence of constant sign; then from property 2—a) and the bound on δ,

$$\mathbf{p} - \delta\mathbf{r}_{N+1} = \sum_{j=1}^{N} \eta_j \mathbf{r}_{a_j} \text{ where } |\eta_j| \leq 1 \ (j = 1, 2, \cdots, n).$$

[3] By definition, a point $\mathbf{p}$ is a boundary point of a set S if every deleted neighborhood of $\mathbf{p}$ contains a point $\mathbf{p}' \in S$ and a point $\mathbf{p}'' \notin S$ [14].

Therefore $\boldsymbol{p}-\delta\boldsymbol{r}_{N+1}\in R_N'$, which is identical with $C(V_N)$ by the induction assumption, thus $\boldsymbol{p}-\delta\boldsymbol{r}_{N+1}\in C(V_N)$. To establish that $\boldsymbol{p}+\delta\boldsymbol{r}_{N+1}\notin C(V_N)$, note that whichever way the a_i's are selected in property 2—a), the n scalars ξ_{a_i} have alternating signs. Since $\xi_{a_n}>0$ and since the ϵ sequence of $\boldsymbol{p}$ has only n subsequences of constant sign, it is easy to see that it is impossible to reduce $\boldsymbol{p}+\delta\boldsymbol{r}_{N+1}$ to the form

$$\sum_{i=1}^{N}\zeta_i\boldsymbol{r}_i \quad \text{with} \quad |\zeta_i| \leq 1 \text{ for } i = 1, 2, \cdots, N.$$

Therefore $\boldsymbol{p}+\delta\boldsymbol{r}_{N+1}\notin R_N'$ and, by the induction assumption, $\boldsymbol{p}+\delta\boldsymbol{r}_{N+1}\notin C(V_N)$.

Statements 2 and 3 above are justified in a similar manner. It is of interest to note that the derivation above shows that every point of $V_N{}^+$, $V_N{}^-$ or $V_N{}^b$ is a boundary point.

From the above statements it follows immediately that the faces of R_N' are subsets of $(n-1)$-dimensional vector spaces spanned by $(n-1)$ vectors of the set $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N\}$.

Basic Property: All edges of the polyhedron R_N' are parallel to one of the vectors $2\boldsymbol{r}_1, 2\boldsymbol{r}_2, \cdots, 2\boldsymbol{r}_N$.

From the descriptions (13) of R_i' for $i\leq n$, this is intuitively clear. For larger N, it seems less obvious. Now, V_N is obtained from V_{N-1} by translating all points of $V_{N-1}{}^+$ and $V_{N-1}{}^b$ by $+\boldsymbol{r}_N$ and all points of $V_{N-1}{}^-$ and $V_{N-1}{}^b$ by $-\boldsymbol{r}_N$. Therefore, all edges of R_{N-1}' undergo a translation before becoming an edge of R_N', and to each point of $V_{N-1}{}^b$ is associated an edge of R_N' which is parallel to $2\boldsymbol{r}_N$. Therefore the above statement follows by induction.

This reasoning leads to an important conclusion: the vertices of R_N' (*i.e.*, the points of V_N) are obtained from the vertices of R_{N-1}' (*i.e.*, the points of V_{N-1}) by: 1) translating all vertices of $V_{N-1}{}^+$ and $V_{N-1}{}^b$ by $\boldsymbol{r}_N$, and 2) translating all vertices of $V_{N-1}{}^-$ and $V_{N-1}{}^b$ by $-\boldsymbol{r}_N$.

E. R_N and Canonical Representation of Points in R_N

By definition, R_N is the set of all points $\boldsymbol{\gamma}$ in the state space from which the origin can be reached in N sampling periods and *no less.*

It is obvious that $R_N=R_N'-R_{N-1}'$.

The question is then: given a point $\boldsymbol{p}\in R_N'$, how is it possible to determine whether $\boldsymbol{p}$ is in R_N or in R_{N-1}'? Let us consider an example in which $N=n+2$ and the point is $\boldsymbol{p}=\boldsymbol{r}_1$. By (12), $\boldsymbol{p}\in R_1'$ and also in R_{n+2}'. Now since the vectors $\boldsymbol{r}_2, \cdots, \boldsymbol{r}_{n+2}$ are linearly dependent, there exist scalars η_i such that

$$\sum_{i=2}^{n+2}\eta_i\boldsymbol{r}_i = 0, \quad \text{with} \quad |\eta_i| \leq 1.$$

Hence with $\eta_1=1$, $\boldsymbol{p}$ has also the representation

$$\boldsymbol{p} = \sum_{i=1}^{n+2}\eta_i\boldsymbol{r}_i.$$

One might erroneously be led to believe from this last relation that $\boldsymbol{p}\in R_{n+2}$. To resolve this ambiguity it is necessary to examine the question of uniqueness of representation. To start with, the following definition is introduced.

Definition: Given a point $\boldsymbol{p}$ in R_N' we have

$$\boldsymbol{p} = \sum_{i=1}^{N}\xi_i\boldsymbol{r}_i \quad \text{where} \quad |\xi_i| \leq 1 \ (i = 1, 2, \cdots, N).$$

To the sequence $\{\xi_i\}$ is associated a sequence $\{\xi_i{}^*\}$ of $+1$'s and -1's by the following rule: 1) $\xi_i=\xi_i{}^*$ for all the i's for which $|\xi_i|=1$. 2) For all others, $\xi_i{}^*$ is assigned the value $+1$ or -1; the choice is made so that the resulting sequence $\{\xi_i{}^*\}$ has the largest number of sign variations. The number of sign variations of the sequence $\{\xi_i{}^*\}$ is called *the maximal number of sign variations of the sequence* $\{\xi_i\}$; it is abbreviated by $\mathcal{V}\{\xi_i\}$. Observe that if $\xi_i{}^*=-1$, then $\xi_i<1$ and if $\xi_i{}^*=1$, then $\xi_i>-1$.

Theorem 2: Let

$$\boldsymbol{p} = \sum_{i=1}^{N}\xi_i\boldsymbol{r}_i, \qquad |\xi_i| \leq 1, \ (i = 1, 2, \cdots, N);$$

hence $\boldsymbol{p}\in R_N'$. The point $\boldsymbol{p}$ has a unique representation of the above form if and only if $\mathcal{V}\{\xi_i\}\leq n-1$.

Theorem 3: Let

$$\boldsymbol{p} = \sum_{i=1}^{N}\xi_i\boldsymbol{r}_i \qquad |\xi_i| \leq 1, \ i = 1, 2, \cdots, N;$$

let $N>n$. Then $\boldsymbol{p}$ is a boundary point of R_N' if and only if the representation of $\boldsymbol{p}$ is unique.

Corollary 1: $\boldsymbol{p}$ is a boundary point of R_N' for $N>n$ if and only if $\mathcal{V}(\xi_i)\leq n-1$.

Corollary 2: For $N>n$, any point of V_N, *i.e.*, any vertex of R_N', has a representation of the following form:

$$\boldsymbol{v} = \sum_{i=1}^{N}\beta_i\boldsymbol{r}_i$$

where

$$|\beta_i| = 1, \ 1 \leq i \leq N, \quad \text{and} \quad \mathcal{V}\{\beta_i\} \leq n-1.$$

The proofs of Theorems 2 and 3 are given in Appendix III.

The consequences of these facts is that given a representation

$$\boldsymbol{p} = \sum_{i=1}^{N}\eta_i\boldsymbol{r}_i$$

one can immediately determine whether or not this point is a boundary point of R_N'. With these tools it is now possible to derive a unique representation of points in R_N. Roughly speaking, a point in R_N is represented as a vector sum of a boundary point of R_{N-1}' and $\delta\boldsymbol{r}_N$ where $|\delta|\leq 1$. To be more precise, let us define:

1) S_{N-1}^{+} that part of the boundary of R_{N-1}' which is made up of the faces whose vertices are in V_{N-1}^{+} or V_{N-1}^{b}.

2) S_{N-1}^{-} that part of the boundary of R_{N-1}' which is made up of the faces whose vertices are in V_{N-1}^{-} or V_{N-1}^{b}.

From the above, if $\gamma \in R_N'$ but $\gamma \notin R_{N-1}'$ then either

$$\gamma = \gamma^{+} + \delta r_N \quad \text{where} \quad \begin{cases} \gamma^{+} \in S_{N-1}^{+} \\ 0 < \delta \leq 1 \end{cases} \tag{17}$$

or

$$\gamma = \gamma^{-} - \delta r_N \quad \text{where} \quad \begin{cases} \gamma^{-} \in S_{N-1}^{-} \\ 0 < \delta \leq 1. \end{cases} \tag{18}$$

Clearly such representation is unique, in the sense that it is the only representation of a point of R_N of the form

$$\gamma = \sum_{i=1}^{N-1} \xi_i r_i + \eta r_N \quad \text{where } \mathcal{U}\{\xi_1, \xi_2, \cdots, \xi_{N-1}, \eta\} \leq n-1.$$

Eqs. (17) and (18) give the *canonical representation of a point in* R_N. The canonical representation will be the basis of the optimal control that will bring the system to equilibrium in the minimum number of sampling periods.

IV. The Critical Surface

As a final preliminary step to obtaining the optimal control, it is necessary to construct a hypersurface in state space. This hypersurface is somewhat analogous to the switching surface for continuous systems. In order to simplify the description of the critical surface the case $n=3$ will be considered. Outlined in Appendix IV is the argument necessary for the general case.

First define a set of parallelograms: for any integers i and j, such that $j>i>1$ the point γ belongs to the parallelogram $\mathcal{P}_{ij}^{+}$ if

$$\gamma = \sum_{\alpha=1}^{j} \xi_\alpha r_\alpha$$

where $\xi_1=0$, $\xi_2=\xi_3=\cdots=\xi_{i-1}=-1$, $|\xi_i|\leq 1$, $\xi_{i+1}=\xi_{i+2}=\cdots=\xi_{j-1}=1$ and $0\leq\xi_j\leq 1$. Pictorially this requirement is illustrated by Fig. 4(a). Similarly, $\mathcal{P}_{ij}^{-}$ is defined in the same way except that the signs of the ξ_α's are reversed. See Fig. 4(b).

The *critical surface* (*CS*) is the surface consisting of the collection of all the parallelograms $\mathcal{P}_{ij}^{+}$ and $\mathcal{P}_{ij}^{-}$.

The critical surface is shown in Fig. 5 for the case $n=3$. For synthesizing the computer we shall make use of the intersection property of the *CS*. Let γ be an arbitrary point. The straight line parallel to r_1 and going through γ intersects the *CS* at one and only one point. In other words,

$$\gamma = \hat{} + \phi r_1,$$

where $\hat{\gamma}$ is any point of the *CS*, determines the scalar ϕ uniquely.

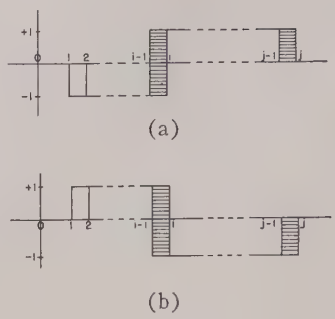


Fig. 4—(a) Representation of the parallelogram $\mathcal{P}_{ij}^{+}$. (b) Representation of the parallelogram $\mathcal{P}_{ij}^{-}$.

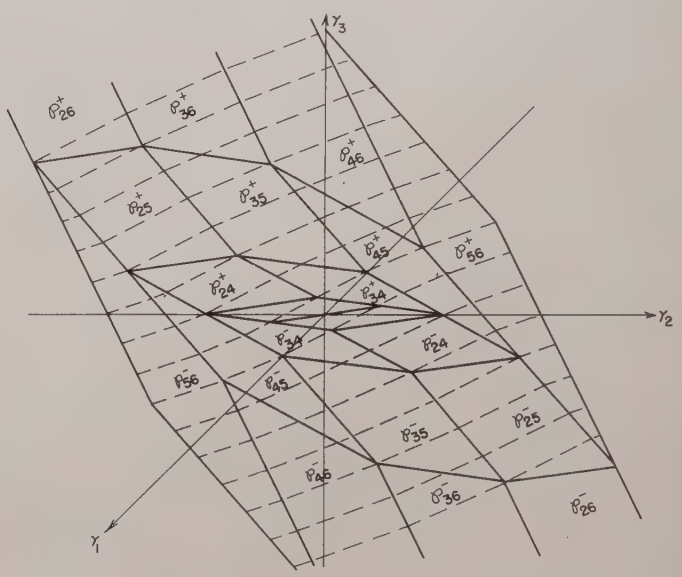


Fig. 5—Critical surface in the neighborhood of the origin for $n=3$.

The uniqueness of ϕ follows from two facts: 1) the *CS* is a connected surface made of the juxtaposition of all the parallelograms $\mathcal{P}_{ij}^{+}$ and $\mathcal{P}_{ij}^{-}$; 2) for all integers i, j, the parallelograms $\mathcal{P}_{ij}^{+}$ and $\mathcal{P}_{ij}^{-}$ are parallel to the plane formed by r_i and r_j; since $j>i>1$, these parallelograms cannot be parallel to r_1 by virtue of property 1.

V. A Computation for an Optimal Control

By *optimal control* is meant the following: given an arbitrary point γ of R_N, an optimal control for γ is a sequence $f_1, f_2, \cdots, f_N$ with $|f_i|\leq 1$, $(i=1, 2, \cdots, N)$ such that

$$\gamma = \sum_{i=1}^{N} f_i r_i.$$

In other words an optimal control for γ is any forcing function satisfying the inequality (2) which brings the point γ to the origin in the minimum time.

In the continuous case it is known that the optimal control is unique [7]. In the discrete case, it has been shown, for the case $n=2$, that except for boundary points of R_N', the optimal control is not unique [11]. The lack of uniqueness of the optimal control may become intuitively obvious from the following step-by-

step characterization of an optimal control: let γ belong to R_N; the forcing function f_1 is the first term of a sequence defining an optimal control if and only if it brings, in one sampling period, the state of the system to *any* point of R_{N-1}. Because of this lack of uniqueness it can only be said that such a computation generates *an* optimal control.

The computation is based on the following theorem.

Theorem 4: Let γ be any point of R_N. Let

$$\gamma = \sum_{i=1}^{N} \eta_i r_i$$

be its canonical representation. If $\gamma = \phi r_1 + \hat{\gamma}$ where $\hat{\gamma} \in CS$ and if

a) $|\phi| \leq 1$ then $f_1 = \phi$ is an optimal control;
b) $\phi > 1$ then $f_1 = 1$ is an optimal control;
c) $\phi < -1$ then $f_1 = -1$ is an optimal control.

Proof:

Case a: $|\phi| \leq 1$. Let

$$\hat{\gamma} = \sum_{a=2}^{N} \xi_a r_a.$$

Then

$$\gamma = \phi r_1 + \sum_{a=2}^{N} \xi_a r_a \tag{19}$$

is the canonical representation of γ because: 1) each coefficient has at most absolute value 1; 2) $\mathcal{V}\{\phi, \xi_2, \xi_3, \cdots, \xi_N\} \leq n-1$. $f = \phi$ is optimal because it follows from (11) that if f_1 is applied now, $\gamma_1 \in R_{N-1}$, *viz.*,

$$\gamma_1 = \Lambda\gamma + \phi d = \sum_{a=2}^{N} \xi_a r_a - 1.$$

Case b: $\phi > 1$. It will be established that $\eta_1 = +1$, consequently $f_1 = +1$ is an optimal control. Consider the case $n = 3$ and let $\mathcal{P}_{ij}^{+}$ be the parallelogram to which $\hat{\gamma}$ belongs. The cross-hatched area in Fig. 6 serves to indicate that $|\xi_i| < 1$ and $0 \leq \xi_j < 1$. The equality (19) is represented by Fig. 6. No generality is lost by reasoning on this particular case, because other cases are treated in a completely similar manner. Let us use property 2 and express r_1 as a linear combination of r_2, r_i and r_j:

$$r_1 = \zeta_2 r_2 + \zeta_i r_i + \zeta_j r_j; \tag{20}$$

then $\zeta_2 > 0$, $\zeta_i < 0$, $\zeta_j > 0$. Using (20) to decrease ϕ the coefficient of r_1 in (19), coefficients of r_2 and r_j will be increased and the coefficient of r_i will be decreased: this is illustrated in Fig. 6 by the vertical arrows. From (19) and (20),

$$\gamma = (\phi - \lambda) r_1 + \sum_{2}^{N} \xi_a r_a + \lambda(\zeta_2 r_2 + \zeta_i r_i + \zeta_j r_j).$$

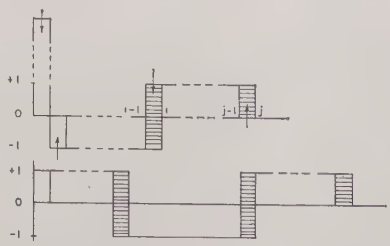


Fig. 6—Transformation of (19), $\phi > 1$, to the canonical representation.

As λ increases, either $\phi - \lambda$ becomes equal to 1 or one of the coefficients of r_2, r_i and r_j becomes equal to 1 in absolute value. In the former case the canonical representation of γ is obtained: thus $\eta_1 = 1$ and the statement is established. In the latter case, let r_a be the vector whose coefficient became first equal to 1 in absolute value; one uses then an expression similar to (20), except that r_a is replaced by r_{a+1}. As one proceeds in this manner one shall eventually reduce ϕ to $+1$ and have

$$\gamma = r_1 + \sum_{i=2}^{N'} \xi_i' r_i, \tag{21}$$

where $\mathcal{V}\{1, \xi_{1'}, \xi_{2'}, \cdots, \xi_{N'}\} \leq n-1$ and $|\xi_i'| \leq 1$. Therefore (21) is a canonical representation and the theorem is established. If $\gamma \in \mathcal{P}_{ij}^{-}$ a similar argument would hold.

Case c is proved in the same manner.

The above theorem gives a means for generating at time $t = 0$ an optimal f_1. Bellman's principle of optimality states [5]: "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

Therefore: An optimal control will be generated if at each sampling instant only the distance, in the direction r_1, between the present state space point and the *CS* is measured and the forcing function is generated according to the rules stated in Theorem 4.

VI. Realization of the Computer ($n = 3$)

Before proceeding with the discussion of the realization of an analog computer for the third-order system, let us introduce the following definition:

Definition: the *critical curve* is that curve obtained by joining successively the vertices defined by

$$\cdots, -\sum_{i=2}^{N} r_i, \cdots, -r_2, +r_2, \cdots, \sum_{i=2}^{N} r_i, \cdots.$$

From this definition and the definition of the parallelograms $\mathcal{P}_{ij}^{+}$ and $\mathcal{P}_{ij}^{-}$, the critical surface in two parts, that which contains the parallelograms $\mathcal{P}_{ij}^{+}$ and that which contains the $\mathcal{P}_{ij}^{-}$. These two parts will be referred to as the positive and negative parts of the *CS*.

From Theorem 4, the computation for the optimal strategy is based on measuring the distance between the

state space point γ and the CS along the direction r_1. Let us therefore consider a set of orthogonal axes Oy_1, y_2, y_3 with the only stipulation that Oy_1 be in the direction of r_1 and the unit of length be chosen equal to the length of r_1. Since any straight line parallel to r_1 intersects the CS at only one point, the CS can be represented as $y_1=\psi(y_2, y_3)$, where $\psi(y_2, y_3)$ is a single-valued piecewise-linear function. To be specific, $\psi(y_2, y_3)$ is the juxtaposition of all the parallelograms $\mathcal{P}_{ij}^{+}$ and $\mathcal{P}_{ij}^{-}$ expressed in terms of the Oy_1, y_2, y_3 coordinate system. The plane supporting $\mathcal{P}_{ij}^{+}$, say π_{ij}^{+}, has an equation of the form

$$y_{1ij}^{+} = a_{ij} + b_{ij}y_2 + c_{ij}y_3, \tag{22}$$

where y_{1ij}^{+} is the first coordinate of the point of π_{ij}^{+} corresponding to the coordinates y_2, y_3. The corresponding expression for the plane π_{ij}^{-} which supports $\mathcal{P}_{ij}^{-}$ has an equation of the form

$$y_{1ij}^{-} = -a_{ij} + b_{ij}y_2 + c_{ij}y_3, \tag{23}$$

where y_{1ij}^{-} is the first coordinate of the point of π_{ij}^{-} corresponding to the coordinates y_2, y_3.

The two following properties of the CS are proven in Appendix V:

Separation Property: For all integers i, j, with $1<i<j$, if π_{ij}^{+} is a plane containing the parallelogram $\mathcal{P}_{ij}^{+}$, then all parallelograms $\mathcal{P}_{kl}^{+}$ for which $k\neq i$, $l\neq j$, and $2\leq k<l$, lie on the same side of the plane π_{ij}^{+}.

Intercept Property: If $y_{1ij}^{+}=a_{ij}+b_{ij}y_2+c_{ij}y_3$ is the equation of the plane π_{ij}^{+} in an orthogonal axes system Oy_1, y_2, y_3 with Oy_1 in the direction of r_1, then the intercept of π_{ij}^{+} with the y_1 axis, *i.e.*, a_{ij}, is negative.

From the symmetry of the CS it follows that the statements of the symmetry property and of the intercept property are still true when *all* superscript $+$'s are replaced by $-$'s.

From the separation property of the CS it follows that if the point (O, y_2, y_3) in the Oy_2, y_3 plane lies within the projection (along Oy_1) of $\mathcal{P}_{kl}^{+}$, then either

(i) $y_{1kl}^{+} > y_{1ij}^{+}$ for all i, j such that $i\neq k, j\neq l$

or

(ii) $y_{1kl}^{+} < y_{1ij}^{+}$

Equivalently, either

$$\text{(i)} \qquad y_{1kl}^{+} = \max_{\substack{i,j \\ i\neq k, j\neq l}} \{y_{ij}^{+}\} \tag{24}$$

or

$$\text{(ii)} \qquad y_{1kl}^{+} = \min_{\substack{i,j \\ i\neq k, j\neq l}} \{y_{ij}^{+}\}.$$

Reference to Fig. 7 shows that (i) is the true relationship. By symmetry,

$$y_{1kl}^{-} = \min_{\substack{i,j \\ i\neq k, j\neq l}} \{y_{ij}^{-}\}. \tag{25}$$

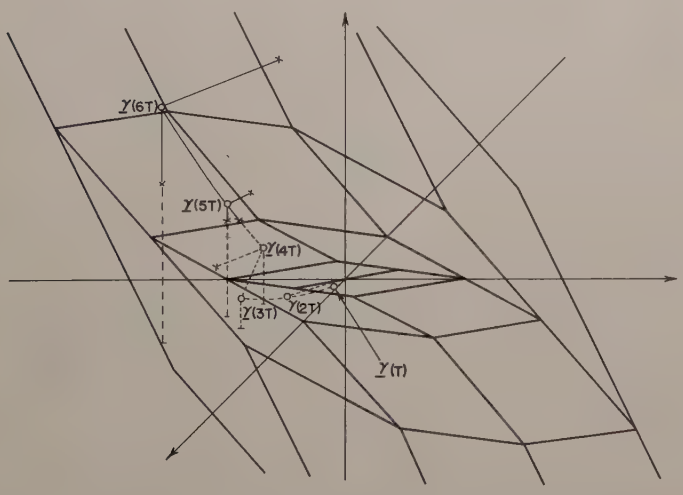


Fig. 7—Geometric interpretation of proposed optimal control.

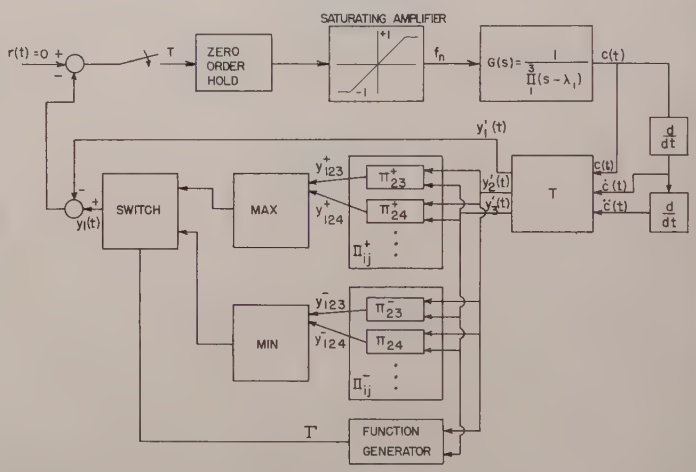


Fig. 8—Block diagram of servomechanism with proposed optimal control computer.

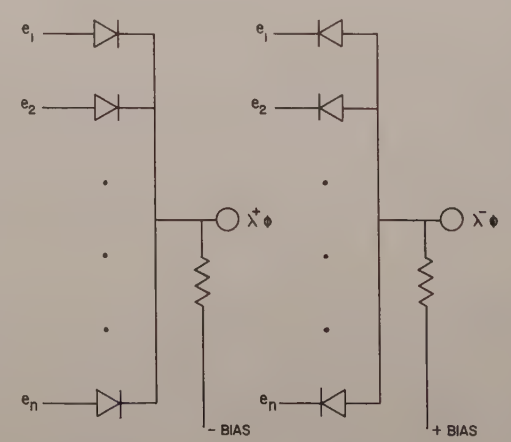


Fig. 9—Physical realization of maximum and minimum network. (a) Maximum network. If $-\text{bias}<\min(e_1, \cdots, e_n)$, then $\lambda\phi^{+}=\max(e_1, \cdots, e_n)$. (b) Minimum network. If $+\text{bias}>\max(e_1 \cdots e_n)$, then $\lambda\phi^{-}=\min(e_1 \cdots e_n)$.

Therefore, if (o, y_2, y_3) belongs to the projection of the positive part of the CS (or the negative part, respectively), then, by (24) [or (25), respectively], a maximization (or minimization, respectively) operation is required. Equivalently, in applying either (24) or (25) to determine the y_1 coordinate of the CS, it is first necessary to determine on which side of Γ the point (o, y_2, y_3) lies, where Γ is the projection of the critical curve on the Oy_2, y_3 plane.

The proposed computer is shown in Fig. 8. Observe that $c(t)$, $\dot{c}(t)$ and $\ddot{c}(t)$ specify the state of the system in the physical coordinates. The block T transforms $c(t)$, $\dot{c}(t)$, $\ddot{c}(t)$ into $y_1'(t)$, $y_2'(t)$, and $y_3'(t)$ which are the coordinates of the state in the Oy_1, y_2, y_3 coordinate system. $y_2'(t)$ and $y_3'(t)$ are inputs to the block Π_{ij}^{+} which contains the adders generating the (22)'s and to the block Π_{ij}^{-} which contains the adders generating the (23)'s. The output of each adder of Π_{ij}^{+} is an input to the maximum network [15] whose output satisfies (24). In a similar manner the output of each adder of Π_{ij}^{-} is an input to the minimum network [15] whose output satisfies (25). The maximum and minimum networks are shown in Fig. 9.

The signal $y_2'(t)$ and $y_3'(t)$ is also an input to the function generator. The function generator generates Γ, the projection of the critical curve on the $y_1=0$ plane. The output of the function generator is positive when $y_2'(t)$, $y_3'(t)$ belongs to a projection of the positive part of the CS and negative otherwise. The ON-OFF signal operates a switch which selects the appropriate y_1 from (24) or (25), thus giving $y_1(t)=\psi[y_2'(t), y_3'(t)]$, where $y_1(t)$ is the ordinate of the point on the CS whose projection on the $y_1=0$ plane is $[o, y_2'(t), y_3'(t)]$. The difference $\phi(t)$ between $y_1'(t)$ and $y_1(t)$ is the distance between the state point and the critical surface. $\phi(t)$ is sampled, passed through a zero hold circuit and then through a saturating amplifier so that the output is f_n for all t in the interval $(nT, (n+1)T)$.

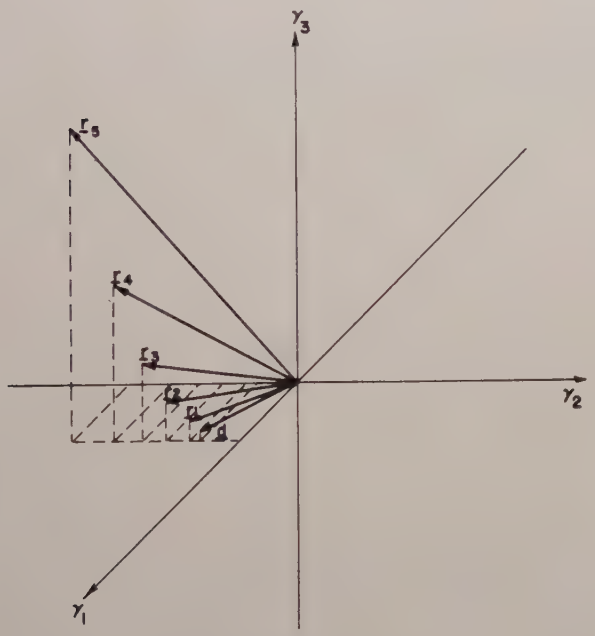


Fig. 10—Representation of the vectors for $n=3$ with a single integral.

VII. Numerical Example

To make the meaning of the proposed optimal strategy concrete, let us consider an example. In Fig. 10 are shown the vectors $\boldsymbol{d}$, $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, $\boldsymbol{r}_3$, $\boldsymbol{r}_4$, and $\boldsymbol{r}_5$ corresponding to a sampling period T and roots 0, $-a$, -1 such that $e^{-T}=0.5$ and $e^{-aT}=0.75$. Figs. 11–13 show regions

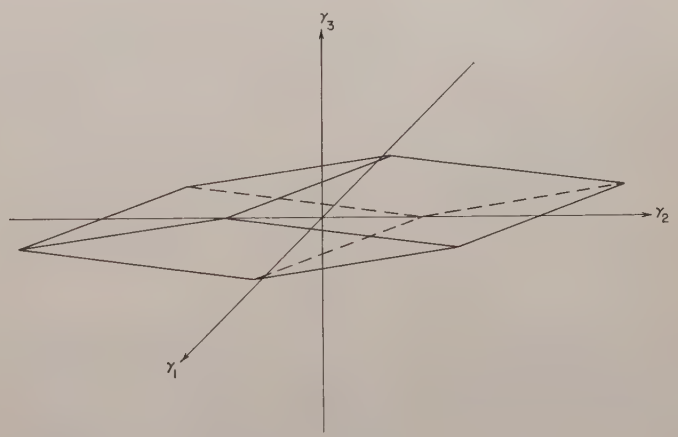


Fig. 12—Set R_3'.

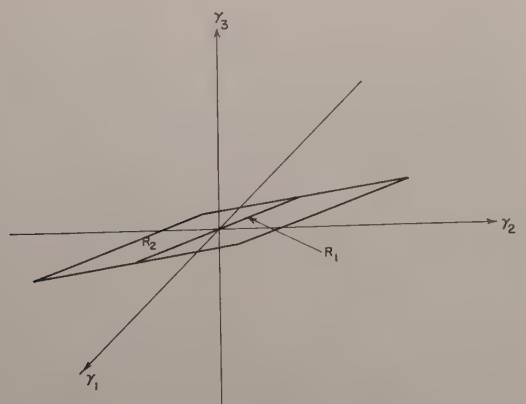


Fig. 11—Sets R_1' and R_2'.

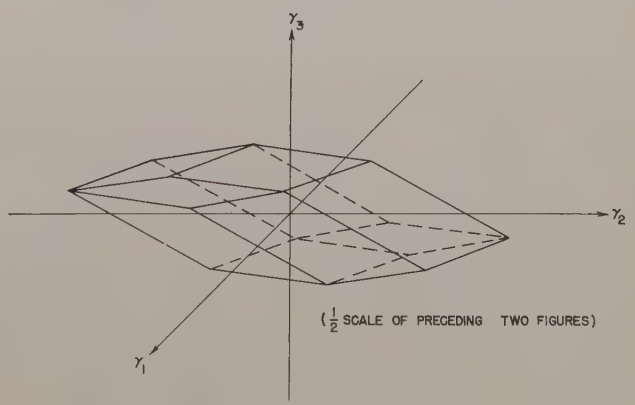


Fig. 13—Set R_4'.

R_1' to R_4' for these numerical values. The monotonic increasing length of r_k for increasing k is clear from the fact that

$$r_k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{0.5k} & 0 \\ 0 & 0 & e^{k} \end{bmatrix} d. \tag{26}$$

In Fig. 5 the CS is shown by utilizing the values of the r_k's of Fig. 10. Contour lines of equal altitude of the CS with respect to the γ_1, γ_2 plane are also indicated.

Let us now take the initial state as

$$\gamma(6T) = r_1 + \tfrac{1}{2}r_2 - r_3 + \tfrac{3}{4}r_4 + r_5 + \tfrac{1}{4}r_6. \tag{27}$$

Since (27) is a canonical representation, then by Theorem 4, if $\gamma(6T) = \phi r_1 + \hat{\gamma}$, where $\hat{\gamma} \in CS$, then $\phi \geq 1$. Fig. 7 shows $\gamma(6T)$ projected onto the CS along a direction parallel to r_1. $\hat{\gamma}$ and ϕ are determined by satisfying

$$[\gamma_3 \text{ of } \hat{\gamma}] = [\gamma_3 \text{ of } (\gamma(NT) - \phi r_1)]. \tag{28}$$

In this case $\phi \geq 1$ so that the optimal control is $+1$, *i.e.*, the coefficient of r_1.

At $t = T$ the state of the system is

$$\gamma(5T) = \tfrac{1}{2}r_1 - r_2 + \tfrac{3}{4}r_3 + r_4 + \tfrac{1}{4}r_5. \tag{29}$$

Reference to Section IV shows that

$$-r_2 + \tfrac{3}{4}r_3 + r_4 + \tfrac{1}{4}r_5 \epsilon \mathcal{P}_{35}^{+}. \tag{30}$$

Applying (28) we find $\phi = 1/2$, which is the coefficient of r_1 for (29). Continuing in the same manner for $t = 2T, \cdots, 6T$ the physical meaning to the proposed optimal control as it relates to the canonical representation of the state is clear.

The proposed optimal control generates the sequence of forcing functions which is identical to the sequence of coefficients of the r_i's of the canonical representation.

VIII. Conclusion

This paper has investigated the minimal-time regulator problem for a saturating sampled-data control system which has a linear plant with real and distinct characteristic roots. An optimal control has been obtained in two stages, first determining the sets R_N' of state space points from which the origin can be reached in N sampling periods or less and second obtaining a unique canonical representation of all points in R_N, the set of state space points from which the origin can be reached in N sampling periods and no less. Finally a block diagram description has been given for an analog computer that generates the proposed optimal control. It is interesting to note that: 1) in the sampled-data case, the optimal control is not unique except for the points on the boundary of the sets R_N'; 2) as T, the sampling period, tends to zero, the length of r_1 goes to zero and the proposed optimal control becomes, in the limit, identical to the usual one for the continuous case: the critical surface becomes the switching surface (this has been shown in detail in [11], for the case $n = 2$); 3) the parallelograms ($\mathcal{P}_{23}^{+}$, $\mathcal{P}_{23}^{-}$) of the CS which contain the origin lie in the plane spanned by r_2 and r_3. Therefore for states close to the origin, our proposed optimal control is identical with the minimal time control for *linear* systems of Kalman ([12], statement (4.12)).

The general techniques used here for discovering and demonstrating rigorously the optimality of our proposed strategy can be applied to more general cases: such as allowing for arbitrary poles for $G(s)$ [17] (both complex and multiple) or as Polak has shown introducing pulse width modulation together with linear or nonlinear plants [18].

Appendix I

A. Proof of Properties 1 and 2

Let $\mu_k = \exp(-\lambda_k T)$; then from (4),

$$\mu_1 > \mu_2 > \cdots > \mu_n \geq 1 > 0. \tag{31}$$

From (10), the components of r_k are

$$r_k = -(\mu_1^k d_1, \mu_2^k d_2, \cdots, \mu_n^k d_n). \tag{32}$$

Therefore the equations that determine the ξ_{α_k} are (we write ξ_k for ξ_{α_k} in the following derivation):

$$\begin{aligned} \mu_1^{\alpha_1}\xi_1 + \mu_1^{\alpha_2}\xi_2 + \cdots + \mu_1^{\alpha_n}\xi_n &= \mu_1^{N+1}, \\ \mu_2^{\alpha_1}\xi_1 + \mu_2^{\alpha_2}\xi_2 + \cdots + \mu_2^{\alpha_n}\xi_n &= \mu_2^{N+1}, \\ &\vdots \\ \mu_n^{\alpha_1}\xi_1 + \mu_n^{\alpha_2}\xi_2 + \cdots + \mu_n^{\alpha_n}\xi_n &= \mu_n^{N+1}. \end{aligned} \tag{33}$$

The determinant Δ of this system of equations may be rewritten as follows:

$$\Delta = (\mu_1, \mu_2, \cdots \mu_n)^{\alpha_1} \begin{vmatrix} 1 & \mu_1^{\beta_2} & \mu_1^{\beta_3} \cdots \mu_1^{\beta_n} \\ \vdots & & \\ 1 & \mu_n^{\beta_2} & \mu_n^{\beta_3} \cdots \mu_n^{\beta_n} \end{vmatrix} \tag{34}$$

where $\beta_i = \alpha_i - \alpha_1$ $(i = 2, 3, \cdots, n)$. Of course $\beta_2 < \beta_3 < \cdots < \beta_n$. Observe that the determinant in the right-hand side of (34) is closely related to a Vandermonde determinant [16]. If we set $\mu_1 = x$ in that determinant, we obtain a function $D(x)$. If we set $x = \mu_2$, or μ_3, or $\cdots$, or μ_n, the determinant vanishes since it would have two identical rows; thus

$$D(x) = (x - \mu_2)(x - \mu_3) \cdots (x - \mu_n)F(x), \tag{35}$$

where $F(x)$ is a polynomial. Suppose we expand the determinant as a function of the elements of the first row: the cofactors are independent of x; therefore $D(x)$ is a polynomial in x having *at most* n coefficients distinct from zero. But from (31) and (35), $D(x)$ has $(n-1)$ positive roots; hence, by Descartes' rule of signs [16], the polynomial $D(x)$ must have alternating signs and cannot have more than $(n-1)$ positive roots. Hence $F(x) \neq 0$ for $x > 0$. Since $\Delta = D(\mu_1)$ and $\mu_1 > 0$, $\Delta \neq 0$. Therefore: 1) the ξ_k's are uniquely determined, and 2) the r_{α_k}'s $(k = 1, 2, \cdots, n)$ are linearly independent. Thus properties 1 and 2—a) are established.

To establish property 2—b), consider the computation of ξ_n. The numerator is identical to Δ except that β_n is replaced by $N+1-\alpha_1$. We still have $\beta_2<\beta_3<\cdots<\beta_{n-1}<N+1-\alpha_1$. Therefore by the above proof the numerator is also positive; hence, sgn $\xi_n=+1$.

To establish property 2—c), we simply observe that the numerator of ξ_i is obtained from Δ by replacing its ith column by col$(\mu_1{}^{N+1}, \mu_2{}^{N+1}, \cdots, \mu_n{}^{N+1})$. After $(n-i)$ permutations of consecutive columns the determinant is of the same form as the above cases, hence sgn $\xi_i=(-1)^{n-1}$.

B. Proof of Property 3

We have to prove that if for some arbitrary integer k,

$$\sum_1^{k+1} \xi_i \boldsymbol{r}_{v_i} = \mathbf{0}, \tag{36}$$

where $v_1<v_2<\cdots<v_{k+1}$, then the sequence $\{\xi_i\}$ has at least n sign variations.

Proof. From property 1, $k\geq n$. From property 2, the above statement is true for $k=n$. To establish the result for any $k\geq n+1$ let us use induction. Therefore we assume that for some $k\geq n+1$, any sum of the form (36) has a corresponding sequence $\{\xi_i\}$ with at least n sign variations and we have to prove that the same property holds for $k+1$. Suppose it would not, *i.e.*, suppose there is a sum

$$\sum_1^{k+1} \delta_i \boldsymbol{r}_{v_i} = \mathbf{0}, \tag{37}$$

for which $\{\delta_i\}$ has $n-1$ sign variations or less. To be specific, let us assume $\{\delta_i\}$ has $n-1$ sign variations,[4] thus $\{\delta_i\}$ is a collection of n subsequences of consecutive δ_i's such that: 1) in each subsequence the δ_i's have the same sign, and 2) the sign alternates from one subsequence to the next. Let us use property 2 to express $\boldsymbol{r}_{v_{k+1}}$ in terms of n of the $\boldsymbol{r}_{v_i}$'s picked in such a way that there is one and only one of them in each subsequence. Let these chosen vectors be $\boldsymbol{r}_{\alpha_j}$ $(j=1, 2, \cdots, n)$. We write

$$\sum_{j=1}^{n} \xi_j \boldsymbol{r}_{\alpha_j} = \boldsymbol{r}_{v_{k+1}} \tag{38}$$

$\{\xi_j\}$ is of alternating sign by property 2. If we add δ_{k+1} times (38) to (37) we get an equation of the form (36) in which $\{\xi_i\}$ has $n-1$ sign variations. This process is illustrated in Fig. 14. In Fig. 14(a), the signs of the δ_i's are shown. In Fig. 14(b), the signs of $\delta_{k+1}\xi_j$ are shown. Since each product $\delta_{k+1}\xi_j$ is of the same sign as the corresponding δ_{α_j}, it is clear that the addition will not change the number of sign variations. This contradicts the induction assumption; therefore the induction is established.

[4] If there are less than $(n-1)$ sign variations the same reasoning applies with a few obvious modifications.

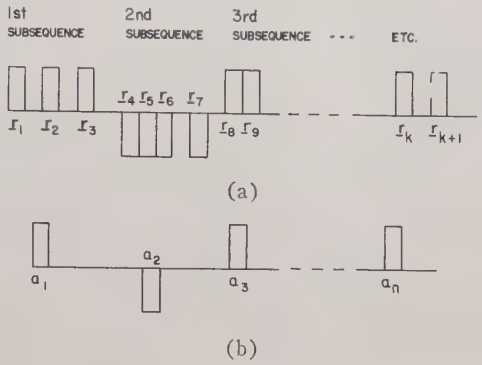


Fig. 14—(a) Signs of the δ_i's. (b) Signs of the $\delta_{k+1}\xi j$'s.

Appendix II

The purpose of this Appendix is to detail the induction proof of the fundamental theorem which asserts that $R_N'=C(V_N)$. Before studying this proof, the reader should be familiar with Section III-D.

1) Consider the set $\hat{V}_{N+1}$ of all points obtained by adding $\boldsymbol{r}_{N+1}$ and subtracting $\boldsymbol{r}_{N+1}$ to each of the points of V_N. Clearly $\hat{V}_{N+1}$ includes V_{N+1}. However, there are points in $\hat{V}_{N+1}$ that do not belong to V_{N+1}; these are obtained by subtracting $\boldsymbol{r}_{N+1}$ from a point of $V_N{}^+$ or by adding $\boldsymbol{r}_{N+1}$ to a point of $V_N{}^-$. In both cases, the result is a point whose ϵ sequence includes n sign variations. These are the only points in $\hat{V}_{N+1}$ that are not included in V_{N+1}.

2) Any point $\boldsymbol{u}$ of R_{N+1}' can be written as

$$\boldsymbol{u} = \sum_{i=1}^{N+1} \xi_i \boldsymbol{r}_i$$

with $|\xi_i|\leq 1$. This sum can be interpreted as a point of R_N', namely

$$\sum_{i=1}^{N} \xi_i \boldsymbol{r}_i,$$

to which is added $\xi_{N+1}\boldsymbol{r}_{N+1}$. Since $R_N'=C(V_N)$ by the induction assumption, $R_{N+1}'=C(\hat{V}_{N+1})$ since $\hat{V}_{N+1}$ is obtained by adding $\pm\boldsymbol{r}_{N+1}$ to all the points of V_N. Therefore the induction will be established if we prove that $C(\hat{V}_{N+1})=C(V_{N+1})$. To establish this fact we need to prove that any point $\boldsymbol{p}$ of $V_N{}^+$ is such that $\boldsymbol{p}-\boldsymbol{r}_{N+1}$ is an interior point of $C(V_{N+1})$ and that any point $\boldsymbol{q}$ of $V_N{}^-$ is such that $\boldsymbol{q}+\boldsymbol{r}_{N+1}$ is an interior point of $C(V_{N+1})$.

3) If $\boldsymbol{p}$ is a point of $V_N{}^+$, it has been shown that it is a boundary point of $C(V_N)$ and by definition, $\boldsymbol{p}-\delta\boldsymbol{r}_{N+1}$ is interior to $C(V_N)$ and $\boldsymbol{p}+\delta\boldsymbol{r}_{N+1}$ is exterior to $C(V_N)$. (See Fig. 15.) Note that $\boldsymbol{p}+\boldsymbol{r}_{N+1}$ is a point of V_{N+1}. Consider the semi-axis $\boldsymbol{p}-\lambda\boldsymbol{r}_{N+1}$, $(\lambda>0)$; since $C(V_N)$ is bounded and convex, it will intersect the boundary of $C(V_N)$ at only one point $\boldsymbol{x}=\boldsymbol{p}-\lambda'\boldsymbol{r}_{N+1}$. Clearly $\lambda'>\delta>0$. This boundary point $\boldsymbol{x}$ lies on a face of $C(V_N)$; hence can be written as follows:

$$\boldsymbol{x} = \sum_{i=1}^{n} \lambda_i \boldsymbol{s}_i$$

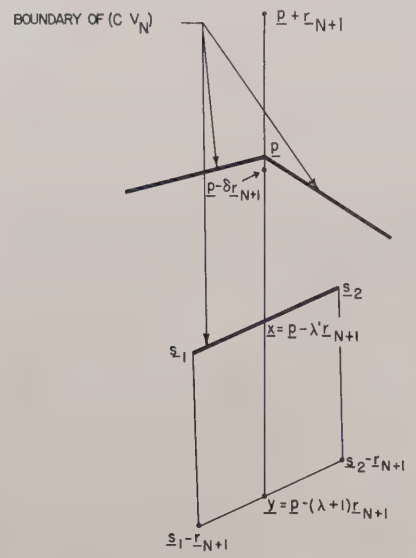


Fig. 15—Geometric basis for proof of fundamental theorem.

where the s_i are vertices of $C(V_N)$, $\lambda_i \geq 0$, and

$$\sum_{i=1}^{n} \lambda_i = 1.$$

We shall prove, below, that none of the s_i may belong to V_N^+. Therefore all points $s_i - r_{N+1} \in V_{N+1}$. In particular, the point y defined by

$$y = x - r_{N+1} = \sum_{i=1}^{N} \lambda_i (s_i - r_{N+1}) = p - (\lambda' + 1) r_{N+1}$$

is a boundary point of $C(V_{N+1})$ and is on the axis $p - \lambda r_{N+1}$. Since $p + r_{N+1}$ and $p - (\lambda'+1) r_{N+1}$ are points of $C(V_{N+1})$ and since $\lambda' > 0$, $p - r_{N+1}$ is an interior point of $C(V_{N+1})$. The same proof, with slight modifications, applies to $q + r_{N+1}$. Thus, $C(\hat{V}_{N+1}) = C(V_{N+1})$ which was shown to be equal to R_{N+1}.

4) To complete the proof we have to show that none of the points $s_1, s_2, \cdots, s_n$ of V_N may belong to V_N^+. Suppose a certain number of them (including, possibly, all of them) are in V_N^+: more precisely, let $s_{i_1}, s_{i_2}, \cdots, s_{i_k}$ belong to V_N^+. Therefore $s_{i_j} - \delta r_{N+1} \in C(V_N)$ for $j = 1, 2, \cdots, k$. This implies a contradiction. Let $u \cdot z = c$ be the equation of the hyperplane containing $s_1, s_2, \cdots, s_n$; *i.e.*, $u \cdot s_i = c$. This hyperplane is a face of the convex polyhedron $C(V_N)$; therefore it is a supporting hyperplane of $C(V_N)$: that is, the sign of u can be chosen so that $u \cdot y \geq c$ for all $y \in C(V_N)$. Now $p = x + \lambda' r_{N+1} \in C(V_N)$ and is not in the hyperplane $u \cdot z = c$, hence, since $\lambda' > 0$, $u \cdot r_{N+1} > 0$. On the other hand, it is alleged that $s_{i_j} - \delta r_{N+1} \in C(V_N)$, hence, since $\delta > 0$, $-u \cdot r_{N+1} \geq 0$, which contradicts the previous inequality. Therefore none of the points $s_1, s_2, \cdots, s_n$ may belong to V_N^+. This concludes the proof that $C(V_N) = R_N'$.

Appendix III

Proof of Theorem 2

We prove first that if the representation is unique, then $\mathcal{V}\{\xi_i\} \leq n-1$. Let us prove this statement by contradiction: *i.e.*, we assume $\mathcal{V}\{\xi_i\} \geq n$. Therefore in the associated sequence $\{\xi_i^*\}$ there are at least n sign variations. Thus it is possible to find indexes $\alpha_1, \alpha_2, \cdots, \alpha_n$ with $\alpha_1 < \alpha_2 < \cdots < \alpha_n$ such that the sequence $\xi_{\alpha 1}^*, \xi_{\alpha_2}^*, \cdots, \xi_{\alpha_n}^*, \xi_N^*$ is of alternating sign. For definiteness let $\xi_N^* = -1$ which implies $\xi_N < 1$. (If $\xi_N^* = +1$ the following reasoning needs a few trivial modifications.) Now from property 2, if

$$\sum_{i=1}^{n} \eta_{\alpha_i} r_{\alpha_i} + r_N = 0,$$

the η_{α_i}'s alternate in sign and $\operatorname{sgn}(\eta_{\alpha_n}) = -1$.

Therefore, using the observation relating the ξ_i^* and ξ_i, there exists an $\epsilon > 0$ sufficiently small such that in the relation

$$p = \sum_{i=1}^{N} \xi_i r_i + \epsilon \sum_{i=1}^{N} \eta_{\alpha_i} r_{\alpha_i} - \epsilon r_N \equiv \sum_{i=1}^{N} \zeta_i r_i$$

(where ζ_i is the resulting factor of r_i), we have

$$|\zeta_i| \leq 1 \qquad (i = 1, 2, \cdots, n).$$

Consequently if $\mathcal{V}\{\xi_i\} \geq n$, the representation is not unique. Therefore we have established that if the representation is unique, $\mathcal{V}\{\xi_i\} \leq n-1$.

We prove next that $\mathcal{V}\{\xi_i\} \leq n-1$ implies that the representation is unique. Again we use contradiction: suppose the representation is not unique, *i.e.*,

$$p = \sum_{i=1}^{N} \xi_i r_i = \sum_{i=1}^{N} \eta_i r_i, \qquad |\xi_i| \leq 1, \ |\eta_i| \leq 1$$

for $i = 1, 2, \cdots, n$;

then,

$$\sum_{i=1}^{N} (\eta_i - \xi_i) r_i = 0,$$

where at least $n+1$ of the $(\eta_i - \xi_i)$ are different from zero. Furthermore, by property 3, $\{\eta_i - \xi_i\}$ has at least n sign variations.

To be specific, let the first nonzero $\eta_i - \xi_i$ be positive. Since for all $0 \leq \lambda \leq 1$, $|\xi_i + \lambda(\eta_i - \xi_i)| \leq 1$, it follows that there are (at least) $n+1$ ξ_i's, say the ξ_{α_j}'s, with the property that $\xi_{\alpha_1} < 1$, $\xi_{\alpha_2} > -1$, $\xi_{\alpha_3} < 1, \cdots$, and $\alpha_1 < \alpha_2 < \cdots < \alpha_{n+1}$. Therefore $\mathcal{V}\{\xi_i\} \geq n$ because the inequalities just obtained show that we can choose $\xi_{\alpha_1}^* = -1$, $\xi_{\alpha_2}^* = +1$, $\xi_{\alpha_3}^* = -1, \cdots$, and these ξ_i^* would already by themselves constitute n sign variations. Therefore $\mathcal{V}(\xi_i) \geq n$. In other words, if the representation is not unique then $\mathcal{V}(\xi_i) \geq n$. Therefore, if $\mathcal{V}(\xi_i) \leq n-1$ the representation must be unique. This concludes the proof.

Proof of Theorem 3

We prove the uniqueness of representation implies that p is on the boundary. First, we exhibit a point

$\boldsymbol{p}' \in R_N'$ which is arbitrarily close to $\boldsymbol{p}$. For any $\epsilon > 0$, arbitrarily small, pick $\boldsymbol{p}' = \sum \xi_i' \boldsymbol{r}_i$ where for $i = 1, 2, \cdots, n$,

$$\xi_i' = \begin{cases} \xi_i - \epsilon & \text{if } \xi_i > \epsilon \\ \xi_i/2 & \text{if } -\epsilon \leq \xi_i \leq \epsilon \\ \xi_i + \epsilon & \text{if } \xi_i < -\epsilon. \end{cases}$$

Next we exhibit a point $\boldsymbol{p}'' \notin R_N'$ which is arbitrarily close to $\boldsymbol{p}$. Take $\boldsymbol{p}'' = \boldsymbol{p} + \epsilon(\operatorname{sgn} \xi_N)\boldsymbol{r}_{N+1}$ with $\epsilon > 0$. Clearly for ϵ arbitrarily small $\boldsymbol{p}''$ is arbitrarily close to $\boldsymbol{p}$. Next, the maximum number of sign variations of the sequence $\{\xi_1, \xi_{2_k} \cdots, \xi_N, \epsilon(\operatorname{sgn} \xi_N)\}$ is $\mathcal{V}\{\xi_i\}$. From the assumed uniqueness and Theorem 2, $\mathcal{V}\{\xi_i\} \leq n-1$. Hence the new sequence has also a maximum number of sign variations $\leq n-1$. This fact together with properties 2 and 3 does not allow $\boldsymbol{p}''$ to be reduced to the form

$$\sum_1^N \eta_i \boldsymbol{r}_i$$

with $|\eta_i| \leq 1$, by writing $\boldsymbol{r}_{N+1}$ as a linear combination of n or more of the preceding $\boldsymbol{r}_i$: it is easy to see that any such attempt would result in some of the ξ_i becoming larger than one in absolute value. Hence $\boldsymbol{p}'' \notin R_N'$ and $\boldsymbol{p}$ is a boundary point of R_N'.

We prove next that if $\boldsymbol{p}$ is on the boundary, then the representation is unique.

Since $\boldsymbol{p}$ is a boundary point and since each face of R_N' is a subset of any of the $(n-1)$ dimensional subspaces spanned by $(n-1)$ vectors of the set $(\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N)$ and since $\boldsymbol{r}_{N+1}$ is outside any such subspace (property 1), at least one of the points $\boldsymbol{p} \pm \epsilon(\operatorname{sgn} \xi_N)\boldsymbol{r}_{N+1}$ (for $\epsilon > 0$ arbitrarily small) must be outside R_N'. The maximum number of sign variations of $\{\xi_1, \xi_2, \cdots, \xi_N, \epsilon(\operatorname{sgn} \xi_N)\}$ is equal to $\mathcal{V}\{\xi_i\}$. Now, by the previous reasoning, it follows that if $\mathcal{V}\{\xi_i\} \geq n$, for ϵ sufficiently small, $\boldsymbol{p} + \epsilon(\operatorname{sgn} \xi_N)\boldsymbol{r}_{N+1}$ and $\boldsymbol{p} - \epsilon(\operatorname{sgn} \xi_N)\boldsymbol{r}_{N+1}$ can be written in the form $\sum \eta_i \boldsymbol{r}_i$, with $|\eta_i| \leq 1$, $i = 1, 2, \cdots, N$. This would mean that both points are in R_N'. Since this is not the case, $\mathcal{V}\{\xi_i\} \leq n-1$ and the representation of $\boldsymbol{p}$ is unique. This concludes the proof.

Appendix IV

The Critical Hypersurface

The purpose of this appendix is to generalize the concept of critical surface in n-dimensional state space.

1) Consider all $(n-1)$ dimensional parallelepipeds defined as follows:

Let the integers $i_1, i_2, \cdots, i_{n-1}$ satisfy the condition $i_1 < i_2 < \cdots < i_{n-1}$. The parallelepiped $\mathcal{P}^+_{i_1 i_2 \cdots i_{n-1}}$ ($\mathcal{P}^-_{i_1 i_2 \cdots i_{n-1}}$, respectively) is the set of all points

$$\boldsymbol{\gamma} = \sum_{j=1}^{i_{n-1}} \xi_j \boldsymbol{r}_j, \tag{39}$$

where

$\xi_1 = 0$,

$\xi_{i_1}, \xi_{i_2}, \cdots, \xi_{i_{n-2}}$ may take any value in the interval $[-1, +1]$,

$\xi_{n-1} \in [0, +1]$ ($\xi_{n-1} \in [-1, 0]$, respectively), all other ξ_j take values ± 1 so that $\mathcal{V}(\xi_j) \leq n-2$.

2) The CS is the collection of all such parallelepipeds.

3) *Intersection theorem.* Given any point $\boldsymbol{\gamma}$ of the state space, there exists one and only one scalar ϕ such that

$$\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}} + \phi \boldsymbol{r}_1 \tag{40}$$

where $\hat{\boldsymbol{\gamma}} \in CS$.

a) Let us first establish existence. Let the point $\boldsymbol{\gamma}$ belong to R_N and let it have the canonical representation

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}' + \delta \boldsymbol{r}_N, \tag{41}$$

where

$$\boldsymbol{\gamma}' = \sum_1^{N-1} \eta_i \boldsymbol{r}_i$$

is a boundary point of R_{N-1}'. Observe that $\mathcal{V}\{\eta_1, \eta_2, \cdots, \eta_{N-1}, \delta\} \leq n-1$ from the definition of a canonical representation. If $\mathcal{V}\{\eta_2, \eta_3, \cdots, \eta_{N-1}, \delta\} \leq n-2$, $\boldsymbol{\gamma} - \eta_1 \boldsymbol{r}_1$ is a point of the critical surface, and a suitable value of ϕ is η_1.

If $\mathcal{V}\{\eta_2, \eta_3, \cdots, \eta_{N-1}, \delta\} = n-1$, there exist $n-1$ elements of the above sequence, say η_{α_1}, such that $\eta_{\alpha_1} < 1$, $\eta_{\alpha_2} > -1$, $\eta_{\alpha_3} < 1, \cdots$. Let us express $\boldsymbol{r}_N$ as a linear combination of $\boldsymbol{r}_1, \boldsymbol{r}_{\alpha_1}, \boldsymbol{r}_{\alpha_2}, \boldsymbol{r}_{\alpha_3}, \cdots, \boldsymbol{r}_{\alpha_{n-1}}$. By property 1, the coefficient δ of $\boldsymbol{r}_N$ in (41) can be reduced in absolute value. This process will stop either because δ is reduced to zero, or because one of the components of $\boldsymbol{r}_{\alpha_1}$, $\boldsymbol{r}_{\alpha_2}, \cdots$ becomes equal to unity in absolute value. In the first case we repeat the procedure using $\boldsymbol{r}_{N-1}$. In the second case, either the resulting sequence η_i' has its maximum number of sign variations reduced by at least 1 or it is still $n-1$. In the former alternative, we have

$$\boldsymbol{\gamma} = \eta_1' \boldsymbol{r}_1 + \sum_{i=1}^{N-1} \eta_i' \boldsymbol{r}_i + \delta' \boldsymbol{r}_N,$$

where $\mathcal{V}\{\eta_2', \eta_3', \cdots, \eta_{N-1}', \delta'\} \leq n-2$. Hence $\phi = \eta_1'$.

In the latter case a new set of elements $\eta_{\beta_1}', \eta_{\beta_2}', \cdots, \eta_{\beta_{n-1}}'$ can be used to repeat the process outlined above.

By successive application of this reduction technique, $\boldsymbol{\gamma}$ can be reduced to the form (40).

b) We establish the uniqueness of ϕ. Write (40) more explicitly:

$$\boldsymbol{\gamma} = \phi \boldsymbol{r}_1 + \sum_{i=2}^{M} \xi_i \boldsymbol{r}_i, \tag{42}$$

where $|\xi_i| \leq 1$ $(i = 2, 3, \cdots, M)$ and $\mathcal{V}\{\xi_i\} \leq n-2$. Thus there are at most $n-1$ ξ_i's, say $\xi_{\alpha_1}, \xi_{\alpha_2}, \cdots$, such that $\xi_{\alpha_1} < 1$, $\xi_{\alpha_2} > -1$, $\xi_{\alpha_3} < 1, \cdots$ (or the same condition in

the opposite signs). But from property 2, when r_1 is expressed as a linear combination of $r_{\alpha_1}, \cdots, r_{\alpha_{n-1}}$ and r_k (where k is arbitrary but larger than M) the coefficients of the combination have $n-1$ sign variations. Therefore any attempt in (42) to vary ϕ will introduce a new sequence ξ_i' with at least $n-1$ sign variations.

Appendix V

Proof of Separation and Intercept Properties

The purpose of this Appendix is to prove the separation and intercept properties of Section VI.

Separation Property

Let

$$\mathbf{x}_{ij} = -r_2 - r_3 - \cdots - r_i + r_{i+1} + \cdots + r_j (1 < i < j).$$

The point $\mathbf{x}_{ij}$ is a vertex of $\mathcal{P}_{ij}^+$. The plane Π_{ij}^+ is the plane passing through the points $\mathbf{x}_{ij}$, $\mathbf{x}_{i+1,j}$, $\mathbf{x}_{i,j+1}$. Let $\mathbf{x}_{ij}'$ designate the four-dimensional row vector whose first three components are identical with those of $\mathbf{x}_{ij}$ and whose last component is unity. Let $\mathbf{x}'$ be the corresponding four-vector to an arbitrary three-vector $\mathbf{x}$. The plane Π_{ij}^+ has an equation given by

$$\det(\mathbf{x}', \mathbf{x}_{ij}', \mathbf{x}_{i+1,j}', \mathbf{x}_{i,j+1}') = 0,$$

where the four vectors in the parentheses are the rows of the determinant. Let $\mathbf{y} = (\eta_1, \eta_2, \eta_3)$ be an arbitrary point of $\mathcal{P}_{kl}^+$ where k and l are fixed. Let $\mathbf{y}' = (\eta_1, \eta_2, \eta_3, 1)$. The separation property will be established if we show that

$$\det(\mathbf{y}', \mathbf{x}_{ij}', \mathbf{x}_{i+1,j}', \mathbf{x}_{i,j+1}') > 0 \qquad \text{for all } k, l,$$

because it would imply that for all i, j with $1 < i < j$, all points $\mathbf{y}$ of all $\mathcal{P}_{kl}^+$, where $k \neq i$ or $l \neq j$, be on the same side of Π_{ij}^+.

Let us establish the property in the case $k > i$, $l > j$. Let $k = i+\alpha$, $l = j+\beta$, hence $\alpha > 0$ and $\beta > 0$. We have to show that

$$\det(\mathbf{x}_{i+\alpha,j+\beta}, \mathbf{x}_{ij}, \mathbf{x}_{i+1,j}, \mathbf{x}_{i,j+1}) > 0 \quad \text{for all } \alpha > 0, \beta > 0,$$

because we need only establish the property for the vertices of the parallelograms $\mathcal{P}_{kl}^+$.

Let Δ be the above determinant. Subtract $\mathbf{x}_{ij}$ from each row of Δ, thus the last column has only one nonzero element. Expanding with respect to that row, we get

$$\Delta = -\det(\mathbf{x}_{i+\alpha,j+\beta} - \mathbf{x}_{ij}, \mathbf{x}_{i+1,j} - \mathbf{x}_{ij}, \mathbf{x}_{i,j+1} - \mathbf{x}_{ij}),$$

where the vectors are unprimed since they are now three-vectors. Let us detail the calculation for the case $i+\alpha \leq j$.

$$\Delta = -\det\left(-2\sum_{q=1}^{\alpha} r_{i+q} + \sum_{q=1}^{\beta} r_{j+q}, -2r_{i+1}, r_{j+1}\right),$$

and by the linear property of the determinant,

$$\Delta = 4\sum_{q=1}^{\alpha} \det(r_{i+q}, -r_{i+1}, r_{j+1}) - 2\sum_{q=1}^{\beta} \det(r_{j+q}, -r_{i+1}, r_{j+1}),$$

$$\Delta = 4\sum_{q=1}^{\alpha} \det(r_{i+1}, r_{i+q}, r_{j+1}) + 2\sum_{q=1}^{\beta} \det(r_{i+1}, r_{j+1}, r_{j+q}).$$

Since $i+\alpha \leq j$, then $i+q < j+1$ and each of the determinants in the above sum is of the form

$$\det(r_\alpha, r_\beta, r_\gamma) = d_1 d_2 d_3 \mu_1^{\alpha+1} \mu_2^{\alpha+1} \mu_3^{\alpha+1} \begin{vmatrix} 1 & 1 & 1 \\ \mu_1^{\beta-\alpha} & \mu_2^{\beta-\alpha} & \mu_3^{\beta-\alpha} \\ \mu_1^{\mu-\alpha} & \mu_2^{\mu-\alpha} & \mu_3^{\mu-\alpha} \end{vmatrix},$$

with $\alpha < \beta < \gamma$ and the notations of (8) with $\mu_i = \exp \lambda_i T$. Since $0 < \mu_1 < \mu_2 < \mu_3$, the determinant is a positive Vandermonde determinant. Hence $\Delta > 0$.

The other cases are proven in a similar fashion.

Intercept Property

The parallelogram $\mathcal{P}_{ij}^+$ is defined as

$$\left\{\mathbf{x} : \mathbf{x} = -\sum_{k=1}^{i} r_k + \sum_{i+1}^{j-1} r_k + \eta_1 r_{i+1} + \eta_2 r_j; \quad -1 \leq \eta_1 \leq 1, 0 \leq \eta_2 \leq 1\right\}. \quad (43)$$

The plane Π_{ij}^+ is then

$$\left\{\mathbf{x} : \mathbf{x} = -\sum_{k=2}^{i} r_k + \sum_{i+2}^{j-1} r_k + \lambda_1 r_{i+1} + \lambda_2 r_j; \lambda_1, \lambda_2 \text{ arbitrary.}\right\}$$

To find a_{ij}, the intercept of Π_{ij}^+ with the Oy_1 axis, set $\mathbf{x} = \mu r_1$ in the equation of Π_{ij}^+, thus, with obvious notations,

$$\mu r_1 = \mathbf{a} + \mathbf{b} + \lambda_1 r_{i+1} + \lambda_2 r_j \quad (44)$$

or

$$\mu r_1 - \lambda_1 r_{i+1} - \lambda_2 r_j = \mathbf{a} + \mathbf{b}.$$

Solving for μ, we get

$$\mu = \frac{\det(\mathbf{a} + \mathbf{b}, r_{i+1}, r_j)}{\det(r_1, r_{i+1}, r_j)} = \frac{\det(\mathbf{a}, r_{i+1}, r_j)}{\det(r_1, r_{i+1}, r_j)} + \frac{\det(\mathbf{b}, r_{i+1}, r_j)}{\det(r_1, r_{i+1}, r_j)}.$$

Since

$$\mathbf{a} = -\sum_{k=1}^{i} r_k \quad \text{and} \quad \mathbf{b} = \sum_{i+2}^{j-1} r_k,$$

it follows that each determinant is negative and $\mu < 0$. Therefore $a_{ij} < 0$.

The equation of Π_{ij}^- is identical to (43) except for a change of sign of both summations. Hence in (44), $\mathbf{a}$ and $\mathbf{b}$ change sign, as does μ. Therefore in the equation of Π_{ij}^-, a_{ij} appears with a minus sign.

References

[1] D. C. McDonald, "Nonlinear techniques for improving servo performance," *Proc. NEC*, vol. 6, pp. 400–421; 1950.
[2] A. M. Hopkin, "A phase plane approach to the design of saturating servomechanism," *Trans. AIEE*, vol. 70, pt. 1, pp. 631–639; 1951.
[3] D. Bushaw, "Optimal discontinuous forcing terms" in "Contribution to the Theory of Nonlinear Oscillations," S. Lefshetz, Ed., Princeton University Press, Princeton, N. J., vol. IV, pp. 29–58; 1958.
[4] I. Bogner and L. F. Kazda, "An investigation of the optimum switching criteria for higher order servomechanisms," *Trans. AIEE*, vol. 73, pt. 2, pp. 118–127; July, 1954.
[5] R. Bellman, I. Clicksberg, and O. Gross, "On the bang bang control problem," *Quart. J. Appl. Math.*, vol. 14, no. 1, pp. 11–18; 1956.
[6] C. A. Desoer, "The bang bang control problem treated by variational techniques," *Information and Control*, vol. 2, pp. 333–348; December, 1959.
[7] V. G. Boltyanski, R. V. Gamkrelidze, E. F. Mishchenko, and L. S. Pontryagin, "The maximum principle in the theory of optimal processes control," *Proc. Internatl. Federation of Automatic Control Congress*, Moscow, USSR, vol. 2, pp. 1004–1008, 1960; Butterworth's Scientific Publications, London, England, 1960.
[8] N. N. Krasovskii, "On the theory of optimum control," *Prikl. Mat. Mekh.*, vol. 23, no. 4, pp. 625–639; 1959.
[9] E. B. Lee, "Mathematical aspects of the synthesis of linear minimum response-time controllers," IRE Trans. on Automatic Control, vol. AC-5, pp. 283–289; July, 1960.
[10] R. E. Kalman, "Optimal nonlinear control of saturating systems by intermittent action," 1957 IRE WESCON Convention Record, pt. 4, pp. 130–135.
[11] C. A. Desoer and J. Wing, "An optimal strategy for a saturating sampled data system," IRE Trans. on Automatic Control, vol. 6, pp. 5–15; February, 1961. Also Electronics Research Lab., University of California, Berkeley, Ser. No. 60, Issue No. 262; December 15, 1959.
[12] R. E. Kalman, "On the General Theory of Control Systems," *Proc. 1st Internatl. Congress of Automatic Control*, Moscow, USSR; 1960.
[13] B. Friedman, "Principles and Techniques of Applied Mathematics," John Wiley and Sons, Inc., New York, N. Y.; 1956.
[14] D. Blackwell and M. A. Girshick, "Theory of Games and Statistical Decisions," John Wiley and Sons, Inc., New York, N. Y.; 1954.
[15] T. E. Stern, "Piecewise-linear network analysis, and synthesis," *Proc. Symp. on Nonlinear Circuit Analysis*, Polytech. Inst. Brooklyn, Brooklyn, N. Y., pp. 315–345; 1956.
[16] J. V. Uspensky, "Theory of Equations," McGraw-Hill Book Co., Inc., New York, N. Y.; 1948.
[17] C. A. Desoer and J. Wing, "Minimal Time Regulator Problem for Linear Sampled Data Systems: General Theory," submitted for publication in *J. Franklin Inst.*
[18] E. Polak, "Minimum time control of second order pulse width modulated sampled data systems," *Trans. ASME*, in press.

Theory and Design of High-Order Bang-Bang Control Systems*

M. ATHANASSIADES†, and O. J. M. SMITH†, fellow, ire

***Summary*—In this paper the nonlinear equations which describe the switching hypersurface of an *N*th-order linear time-invariant system, with real negative distinct poles, are developed when the input to the system is restricted in amplitude, either intentionally through the use of a relay or due to saturation of the power element.**

Based on the equations of the switching hypersurfaces, a design procedure is offered which will result in the optimum transient system response for any initial output or error initial conditions in the absence of an input.

Introduction

THERE are many papers,[1–8] both in American and Russian literature concerned with the general theory, existence, and properties of the solution, as well as a multitude of approximate designs, for the so-called "bang-bang" control problem.

From a theoretical viewpoint, the "maximum" principle of Pontriagin[1] provides the general properties of the optimal control function, provided that an optimal control function exists. Furthermore, if the controlled system is linear, then the maximum principle is a necessary property of the optimal control function. Also, related papers,[2,6–8] such as the adjoint system method of Desoer,[2] show that the output of the adjoint system may be used to determine the optimal control function; however, Desoer's paper does not describe in detail the nonlinear transformation which transforms the initial output of the actual system to the initial values of the adjoint system.

On the other hand, many approximate solutions for the second- and third-order cases exist in the litera-

* Received by the PGAC, December 10, 1960; revised manuscript received, March 1, 1961.

† Dept. of Elec. Engrg., University of California, Berkeley, Calif.

[1] L. I. Rozonoer, "L. S. Pontriagin's maximum principle in the theory of optimal systems," *Automation and Remote Control*, vol. 20: October, November, December, 1959. (English translation.)

[2] C. A. Desoer, "The bang-bang servo problem treated by variational techniques," *Inform. and Control*, vol. 2, pp. 333–348; Month, 1959.

[3] E. B. Lee, "Design of Optimum Multivariate Control Systems," ASME Paper No. 60-JAC-5; 1960.

[4] E. B. Lee, "Mathematical aspects of the synthesis of linear minimum response time controllers," IRE Trans. on Automatic Control, vol. AC-5, pp. 283–289; September, 1960.

[5] Yu-Chi-Ho, "Solution Space Approach to Optimal Control Problems," ASME Paper No. 60-JAC-11; 1960.

[6] J. P. LaSalle, "Time optimal control systems," *Proc. Natl. Acad. Sciences*, vol. 45, pp. 573–577; April, 1959.

[7] J. P. LaSalle, "The Bang-Bang Principle," RIAS, Baltimore, Md., Rept. No. 59-5; November, 1959.

[8] R. Bellman, I. Glicksberg, and O. Gross, "On the bang-bang control problem," *Quart. Appl. Math.*, vol. 14, pp. 11–18; 1956.

ture.[9–18] However, the approximations used do not stem from the knowledge of the exact solution of the problem, but they are mostly based on vague assumptions.

The purpose of this paper is to present in detail the various linear and nonlinear transformations on the error variables, such that the system transient response is optimal in the minimum error time sense. This paper deals only with an autonomous system, that is, a system with zero input, but variable initial conditions. The optimal transient response is to be interpreted as the system behavior due to arbitrary error or output initial conditions. These nonlinear transformations are carried out by a nonlinear computer, which may be, in practice, either of digital or of analog nature.

Statement of the Problem

Given a plant with the transfer function

$$G(s) = 1/(s + s_1) \cdots (s + s_N), \tag{1}$$

and the feedback connections shown in Fig. 1; the input to the plant is provided by a relay or maximum-effort-type element, and the input to the relay is the manipulated variable $m(t)$, which is the output of the nonlinear computer. The plant output is subtracted from the input to provide the error $e(t)$.

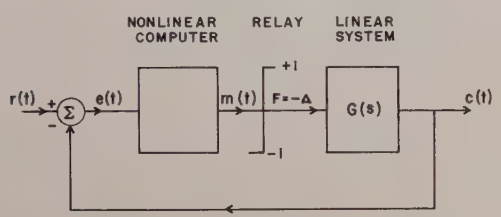


Fig. 1—Block diagram of the system.

[9] I. Bogner, "An Investigation of the Switching Criteria for Higher Order Contactor Servomechanisms," Cook Res. Labs., Chicago, Ill., Interim Rept. No. PR-16-9; 1953.
[10] A. M. Hopkin, "A phase plane approach to the design of saturating servomechanisms," *Trans. AIEE*, vol. 70 (Application and Industry), pp. 631–639; 1950.
[11] D. C. McDonald, "Nonlinear techniques for improving servo performance," *Proc. Natl. Electronics Conf.*, vol. 6, pp. 400–421; 1950.
[12] J. L. Preston, "Nonlinear Control of a Saturating Third Order Servomechanism," Mass. Inst. Tech., Cambridge, M.I.T. Tech. Memo. 6987-TM-14; 1954.
[13] P. Wang, "The Design of Relay Type Control Systems for Random Inputs," Ph.D. dissertation, Univ. of California, Berkeley; June, 1959.
[14] S. S. L. Chang, "Optimum switching criteria for higher order contractor servo with interrupting circuits," *Trans. AIEE*, vol. 74, (Application and Industry), pp. 273–276; November, 1955.
[15] N. C. Walker, "A Predictor Servomechanism," M. S. thesis, Univ. of California, Berkeley; 1953.
[16] L. M. Silva, "Nonlinear Optimization of a Relay Servo," M.S. thesis, Univ. of California, Berkeley; 1953.
[17] R. E. Kuba and L. F. Kazda, "A phase space method for the synthesis of nonlinear servomechanisms," *Trans. AIEE*, vol. 75, (Application and Industry), pp. 282–290; November, 1956.
[18] A. M. Hopkin and M. Iwama, "A Study of a Digitally Programmed Optimum Relay Servomechanism for Nonlinear Control of an Airframe," Electronics Res. Lab., Univ. of California, Berkeley, Rept. No. 60; February, 1957.

It is assumed that the poles of the system are $s_1, \cdots, s_N$ such that

$$0 < s_1 < s_2 < \cdots < s_N. \tag{2}$$

Fig. 2 shows the s-plane plot of the transfer function $G(s)$. The s_j values are positive.

The requirement on the system response is that $c(t) = r(t) = 0$, after a finite transient time which is the minimum possible time.

It is desired to determine the nonlinear computer which, after operating on the error, will generate the manipulated variable $m(t)$ resulting in the satisfaction of the minimum time requirement.

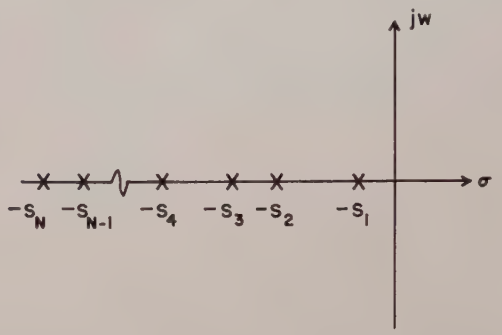


Fig. 2—s-plane pole configuration of the linear plant $G(s)$.

Preliminary Theory

Since the output of the system is to be identical to the input after a finite transient minimum time, then the error and all of its derivatives must be reduced to zero simultaneously in the minimum possible time. Previous investigators have proved the following theorem:

Theorem 1: Given the Nth-order linear time-invariant system, with poles real and nonpositive, connected as in Fig. 1, the error and all of its derivatives may be reduced to zero simultaneously, in the minimum possible time, with at most N trajectories, if the system is controllable, requiring at most $N-1$ switchings (reversals) of the relay, and an initial-on and final-off operation provided that $r(t) = 0$.

The above theorem is a direct consequence of two theorems proved by Bellman.[8] The first of these theorems states that the optimal control function exists and is of the bang-bang type, *i.e.*, $|F(t)| = 1$, and the second theorem shows that the control function needs to change sign at most $N-1$ times.

Let the input to the system be zero. The output of the plant may have any initial value and the error of the system may have any initial value. The problem is to reduce the error and its derivatives to zero in the minimum possible time.

Let the output of the relay be

$$F(t) = -\Delta \tag{3}$$

where $\Delta = +1$ or -1 depending on the sign of the function

$$m(t). \tag{4}$$

Between switchings of the relay, the control function is a constant, hence,

$$C(s) = G(s)(-\Delta/s); \tag{5}$$

therefore, the output of the system satisfies a linear differential equation of the form

$$\frac{d^N c(t)}{dt^N} + a_{N-1}\frac{d^{N-1}c(t)}{dt^{N-1}} + \cdots + a_0 c(t) = -\Delta. \tag{6}$$

Since the input is zero, then,

$$c(t) = -e(t); \dot{c}(t) = -\dot{e}(t); \cdots; c^{(N)}(t) = -e^{(N)}(t). \tag{7}$$

By using (7), (6) becomes

$$\frac{d^N e(t)}{dt^N} + a_{N-1}\frac{d^{N-1}e(t)}{dt^{N-1}} + \cdots + a_1\frac{de(t)}{dt} + a_0 e(t) = \Delta. \tag{8}$$

The Nth-order linear differential equation above may be reduced to a system of N first-order differential equations by the following change of variables

$$y_1 = e; \quad y_2 = \dot{e}; \quad y_3 = \ddot{e}; \cdots; y_N = e^{(N-1)}. \tag{9}$$

Using the above substitution, the Nth-order differential equation is reduced to a system of N first-order linear differential equations,

$$\begin{aligned}
\dot{y}_1 &= y_2 \\
\dot{y}_2 &= y_3 \\
&\cdots \\
\dot{y}_{N-1} &= y_N \\
\dot{y}_N &= -a_0 y_1 - a_1 y_2 - \cdots - a_{N-1} y_{N-1} + \Delta,
\end{aligned} \tag{10}$$

which is of the matrix form

$$\dot{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{f}, \tag{11}$$

where

$$\boldsymbol{y} = (y_1, \cdots, y_N); \qquad \boldsymbol{f} = (0, 0, \cdots, 0, \Delta) \tag{12}$$

are column vectors and the matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{N-1} \end{bmatrix}. \tag{13}$$

Of course, the eigenvalues of the matrix $\boldsymbol{A}$ are the poles of the system.

The matrix differential equation (11) may be reduced to the diagonal form by the following procedure: Let $\boldsymbol{S}$ be the matrix which has as its diagonal the poles of the system, *i.e.*, the eigenvalues of $\boldsymbol{A}$.

$$\boldsymbol{S} = \begin{bmatrix} -s_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -s_2 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & -s_N \end{bmatrix} \tag{14}$$

then, there exists a nonsingular matrix $\boldsymbol{P}$,

$$\boldsymbol{P} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ -s_1 & -s_2 & \cdots & -s_N \\ s_1^2 & s_2^2 & \cdots & s_N^2 \\ \cdots & \cdots & \cdots & \cdots \\ (-s_1)^{N-1} & \cdot & \cdots & (-s_N)^{N-1} \end{bmatrix}, \tag{15}$$

such that

$$\boldsymbol{S} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}, \tag{16}$$

and if

$$\boldsymbol{x} = \boldsymbol{P}^{-1}\boldsymbol{y} \tag{17}$$

$$\boldsymbol{g} = \boldsymbol{P}^{-1}\boldsymbol{f}, \tag{18}$$

then (11) reduces to

$$\dot{\boldsymbol{x}} = \boldsymbol{S}\boldsymbol{x} + \boldsymbol{g}. \tag{19}$$

The transformation of variables (17) is linear, hence, when $\boldsymbol{y} = 0$, then $\boldsymbol{x} = 0$.

Let (b_{ij}) be the elements of the matrix $\boldsymbol{P}^{-1}$,

$$\boldsymbol{P}^{-1} = \begin{bmatrix} b_{11} & \cdots & b_{1N} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ b_{N1} & \cdots & b_{NN} \end{bmatrix}. \tag{20}$$

Note that the columns of the matrix $\boldsymbol{P}$ are eigenvectors of the matrix $\boldsymbol{A}$ and that the rows of the matrix $\boldsymbol{P}^{-1}$ are the reciprocal basis eigenvectors of the matrix $\boldsymbol{A}$.

From (19), one obtains the following differential equations:

$$\begin{aligned}
\dot{x}_1 &= -s_1 x_1 + b_{N1}\Delta \\
\dot{x}_2 &= -s_2 x_2 + b_{N2}\Delta \\
&\cdots \\
\dot{x}_N &= -s_N x_N + b_{NN}\Delta.
\end{aligned} \tag{21}$$

Hence, the variable x_j is the solution of the first-order differential equation

$$\dot{x}_j = -s_j x_j + b_{jN}\Delta \tag{22}$$

which is solved to yield

$$x_j(t) = (x_{j0} - b_{jN}\Delta/s_j)e^{-s_j t} + b_{jN}\Delta/s_j; \quad t > 0 \tag{23}$$

where x_{j0} is the initial value, at $t = 0$, of the variable x_j, and t is positive greater than zero.

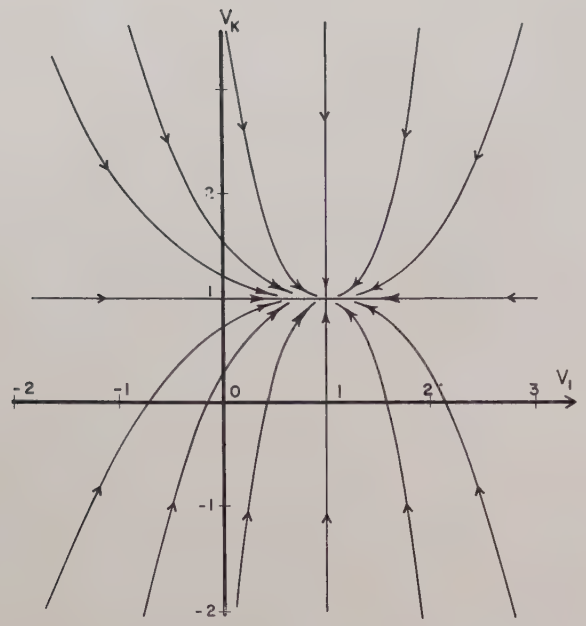


Fig. 3—Plot of positive time trajectories projected in the v_1-v_K phase plane for $\Delta=+1$ and $s_K/s_1=2.5$.

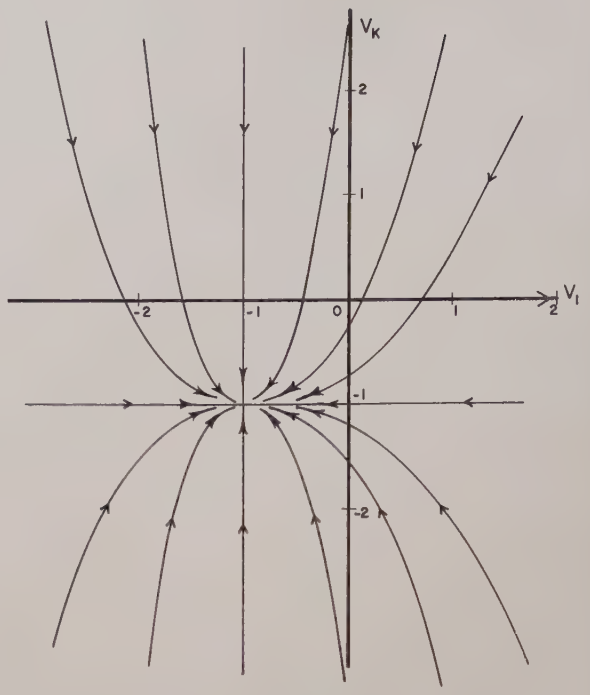


Fig. 4—Plot of positive time trajectories projected in the $v_1=v_K$ phase plane for $\Delta=-1$ and $s_K/s_1=2.5$.

One may apply an additional linear transformation of variables in order to clarify the equations. Let

$$v_j(t) = x_j(t)\frac{s_j}{b_{jN}} \tag{24}$$

using transformation (24), then (23) reduces to

$$v_j(t) = (v_{j0} - \Delta)e^{-s_j t} + \Delta; \quad j = 1, 2, \cdots, N; \quad t > 0. \tag{25}$$

One may solve for the time t_1 in terms of the v_j variables.

$$t = \ln\left(\frac{v_{j0} - \Delta}{v_j(t) - \Delta}\right)^{1/s_j} \tag{26}$$

in particular, for $j=1$

$$t = \ln\left[(v_{10} - \Delta)/(v_1(t) - \Delta)\right]^{1/s_1}; \quad t > 0 \tag{27}$$

substituting (27) into (25) to eliminate time

$$v_k(t) = (v_{k0} - \Delta)\left[(v_1(t) - \Delta)/(v_{10} - \Delta)\right]^{s_k/s_1} + \Delta \qquad k = 2, 3, \cdots, N. \tag{28}$$

Eq. (28) describes the equations of the trajectories in the phase space, *i.e.*, given an initial point and a constant relay output of $-\Delta$, the relation between the trajectory coordinates is uniquely described. Note that (28) describes a curve in N-dimensional phase space. Figs. 3 and 4 show the trajectories of $s_k/s_1=2.5$. The mathematical relation between the v variables and the error variables is given by (9), (17), and (24). Thus, in a physical system one may obtain the v variable as follows:

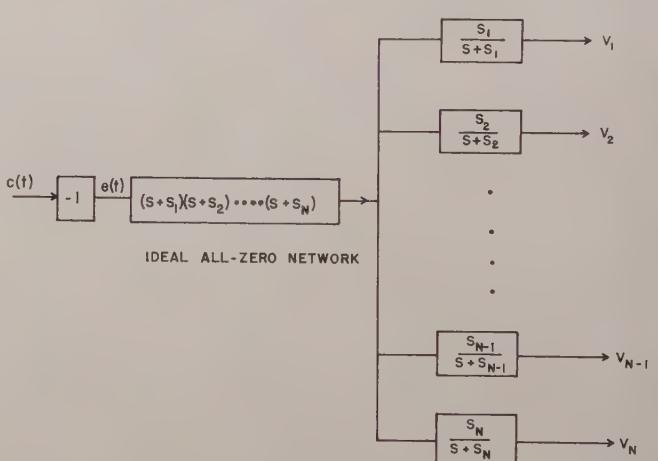


Fig. 5—Network realization of the v variables from the error signal.

1) Differentiate the error $N-1$ times to obtain the y variables.
2) Apply the linear transformation (17) to obtain the x variables after the matrix $\boldsymbol{P}$ is inverted and the (b_{ij}) elements obtained.
3) Apply the linear change of variable (24) to construct the v variables from the x variables. However, in a physical system one may obtain the v variables directly from the error by operating on the error with an all-zero ideal network with transfer function $1/G(s)$, and by operating on the single resulting signal by a collection of lag networks each of the form $s_j/(s+s_j)$ to construct each variable v_j. This procedure is shown in Fig. 5 and the proof is indicated in Appendix I.

Optimal Trajectories and the Concept of Switching Surfaces

There exists a single trajectory, for a given Δ, which is unique and passes through the origin of the phase space. Let the set of points $(V) = (v_1, \cdots, v_N)$, belonging to that trajectory, be described by the subscript $N-1$, (V_{N-1}). This trajectory, referred to as the final trajectory, has the property, that if (V) belongs to (V_{N-1}) then the state point (V) may be brought to the origin without reversing the polarity of the relay output and in the minimum possible time.

In addition, there exists a set of points in the phase space (V_{N-2}) such that if the state point (V) belongs to the set (V_{N-2}) then the terminal point of the trajectory is set (V_{N-1}), hence, the state point (V) belonging to the set (V_{N-2}) may be brought to the origin with only one switching or reversal of the relay and in the minimum possible time.

Since the solutions of differential equations are unique, the polarity of the relay output during the trajectory originating on (V_{N-2}) and terminating on (V_{N-1}) must be opposite to the polarity of the relay output during the trajectory originating on the set (V_{N-1}) and terminating on the origin of the phase space. Furthermore, the path followed from (V_{N-2}) to (V_{N-1}) and from (V_{N-1}) to (0) is optimal, and, although physically obvious, it is a direct consequence of Bellman's "principle of optimality."

Proceeding similarly, one establishes that there exists a set of points (V_{N-3}), such that if (V) belongs to (V_{N-3}) then the state point may be brought to the origin of the phase space with only two switchings or reversals of the relay and in minimum time. Note that trajectories originating on (V_{N-3}) must terminate on (V_{N-2}).

Following the same line of thought, one may establish the following: There exists in the phase space a set of points (V_1) such that if the state point (V) belongs to the set (V_1), the state point may be brought to the origin, in the minimum possible time, with exactly N -2 switchings or reversals of the relay. The set (V_1) is called the switching hypersurface of the system. Clearly, trajectories originating on (V_1) terminate in (V_2). Moreover, trajectories originating on (V_2) terminate on (V_3), and so on. Hence, the set of points (V_1) contains the subset (V_2) which contains (V_3) which contains . . . (V_{N-1}) which contains the origin (0). An arbitrary set of initial conditions, which do not necessarily lie on the switching hypersurface, is the set (V_0).

Technique for Control

Given a state point (V), which describes the present state of the system, one must test to see if (V) belongs to (V_1), *i.e.*, if the state point is on the switching hypersurface. If (V) does not belong to (V_1), then the relay must generate the appropriate polarity such that the resultant trajectory will in finite time terminate on the set (V_1). If (V) belongs to (V_1), then the relay must produce the correct polarity such that the resultant trajectory will terminate on (V_2), and so on. This procedure satisfies the requirement that, given any initial condition, the state point may be brought to the origin with at most $N+1$ relay operations, consisting of a turn on, $N-1$ switchings or reversals, and a turn off of the relay. If the state point reaches the origin, then the error and all of its derivatives are reduced to zero simultaneously and the restrictions on the transient response are satisfied.

The Equation of the Switching Hypersurface

Let Δ^* be the relay output during the final trajectory. Since the origin is the terminal point of the final trajectory, from (28), the final trajectory satisfies the equation

$$0 = \left(\frac{-\Delta^*}{v_{1,N-1} - \Delta^*}\right)^{s_k/s_1} (v_{k,N-1} - \Delta^*) + \Delta^*$$

$$k = 2, 3, \cdots, N. \quad (29)$$

Eq. (29) describes the set (V_{N-1}). It is true only for $t>0$.

The set (V_{N-2}) was defined as the set of all trajectories terminating on the final trajectory (V_{N-1}). During this trajectory the output of the relay is $-\Delta^*$. Therefore,

$$v_{k,N-1} = \left(\frac{v_{1,N-1} + \Delta^*}{v_{1,N-2} + \Delta^*}\right)^{s_k/s_1} (v_{k,N-2} + \Delta^*) - \Delta^*$$

$$k = 2, 3, \cdots, N. \quad (30)$$

$$t > 0$$

The set (V_{N-3}) was defined as the set of all trajectories terminating on the set of trajectories (V_{N-2}). The relay output during these trajectories must be $-(-\Delta^*) = \Delta^*$. Therefore,

$$v_{k,N-2} = \left(\frac{v_{1,N-2} - \Delta^*}{v_{1,N-3} - \Delta^*}\right)^{s_k/s_1} (v_{k,N-3} - \Delta^*) + \Delta^*$$

$$k = 2, 3, \cdots, N. \quad (31)$$

Proceeding as above, back to the second trajectory, one determines the relationship between the (V_1) and (V_2) sets for $t>0$,

$$v_{k,2} = \left(\frac{v_{1,2} - \Delta^*}{v_{1,1} - \Delta^*}\right)^{s_k/s_1} (v_{k,1} - \Delta^*) + \Delta^*$$

$$k = 2, 3, \cdots, N, \text{ for } N \text{ odd.} \quad (32\text{-a})$$

$$v_{k,2} = \left(\frac{v_{1,2} + \Delta^*}{v_{1,1} + \Delta^*}\right)^{s_k/s_1} (v_{k,1} + \Delta^*) - \Delta^*$$

$$k = 2, 3, \cdots, N, \text{ if } N \text{ is even.} \quad (32\text{-b})$$

Eq. (32) describes the switching hypersurface of the system in a parametric form. Note that (29), (30), and

(31) are restrictions imposed on the (V_2), (V_3), $\cdots$, (V_{N-1}) variables. Solving for $v_{k,N-1}$ from (29) and substituting into (30), then solving for $v_{k,N-2}$ from (30) and substituting into (33), etc., the following set of equations is obtained:

$$v_{k,N-1} - \Delta^* = -\Delta^*\left(\frac{v_{1,N-1} - \Delta^*}{-\Delta^*}\right)^{sk/s_1} \tag{33-a}$$

$$v_{k,N-2} + \Delta^* = \left(\frac{v_{1,N-2} + \Delta^*}{v_{1,N-1} + \Delta^*}\right)^{sk/s_1} \cdot\left[2\Delta^* - \Delta^*\left(\frac{v_{1,N-1} - \Delta^*}{-\Delta^*}\right)^{sk/s_1}\right] \tag{33-b}$$

$$v_{k,N-3} - \Delta^* = \left(\frac{v_{1,N-3} - \Delta^*}{v_{1,N-2} - \Delta^*}\right)^{sk/s_1} \cdot\left\{-2\Delta^* + \left(\frac{v_{1,N-2} + \Delta^*}{v_{1,N-1} + \Delta^*}\right)^{sk/s_1} \cdot\left[2\Delta^* - \Delta^*\left(\frac{v_{1,N-1} - \Delta^*}{-\Delta^*}\right)^{sk/s_1}\right]\right\} \tag{33-c}$$

and for N odd

$$v_{k,1} + \Delta^* = \left(\frac{v_{1,1} + \Delta^*}{v_{1,2} + \Delta^*}\right)^{sk/s_1}\left\{2\Delta^* + \left(\frac{v_{1,2} - \Delta^*}{v_{1,3} - \Delta^*}\right)^{sk/s_1} \cdot\left\{-2\Delta^* + \cdots\left(\frac{v_{1,N-3} - \Delta^*}{v_{1,N-2} - \Delta^*}\right)^{sk/s_1} \cdot\left\{-2\Delta^* + \left(\frac{v_{1,N-2} + \Delta^*}{v_{1,N-1} + \Delta^*}\right)^{sk/s_1} \cdot\left\{2\Delta^* - \Delta^*\left(\frac{v_{1,N-1} - \Delta^*}{-\Delta^*}\right)^{sk/s_1}\right\}\right\} \cdots\right\}\right\} \tag{33-n}$$

for $k = 2, 3, \cdots, N$ and for positive time.

Eq. (33-n) indicates that the variables of the switching hypersurface $(V_1) = (v_{1,1}, \cdots, v_{N,1})$ are related by a total of $N-1$ nonlinear algebraic equations which also involve the (unknown) values of the variable v_1 at the instant of the second switching, at the instant of the third switching, . . . , at the instant of the $N-1$ switching of the relay, and also the (unknown) value of $\Delta^* = \pm 1$.

The equations are of the form

$$v_{2,1} = f_2(s_j; \Delta^*; v_{1,1}; v_{1,2}; \cdots; v_{1,N-1}) \tag{34-a}$$

$$v_{3,1} = f_3(s_j; \Delta^*; v_{1,2}; v_{1,2}; \cdots; v_{1,N-1}) \tag{34-b}$$

. . . . . .

$$v_{N-1,1} = f_{N-1}(s_j; \Delta^*; v_{1,1}; v_{1,2}; \cdots; v_{1,N-1}) \tag{34-m}$$

$$v_{N,1} = f_N(s_j; \Delta^*; v_{1,1}; v_{1,2}; \cdots v_{1,N-1})$$

$$\text{for } j = 1, \cdots, N. \tag{35}$$

Thus we have a total of $N-1$ equations involving the relationship between N variables of the switching hypersurface, and, in addition, there are $N-2$ unknowns. In order to obtain the equation of the switching hypersurface, one may solve the $N-2$ equations (34-a) through (34-m) for the unknowns $v_{1,2}, \cdots, v_{1,N-1}$ in terms of the variables $v_{1,1}, v_{2,1}, \cdots, v_{N-1,1}$ and then substitute into (35) the switching surface of the form

$$v_{N-1} = F(s_j; \Delta^*; v_{1,1}, \cdots, v_{N-1,1}) \text{ with } \Delta^* = \pm 1.$$

Thus it is shown how to obtain, at least in principle, the equation of the switching hypersurface, *i.e.*, the equations of the set of points (V_1). The procedure for obtaining the correct value of Δ^* will be developed in the following section.

The Auxiliary Equations for the Switching Surface

Let t_{N-1} be the time required to go from the set (V_{N-1}) to (0).

Let t_{N-2} be the time required to go from the set (V_{N-2}) to (V_{N-1}).

. . . . . .

Let t_{N-R} be the time required to go from the set (V_{N-R}) to (V_{N-R+1}).

. . . . . .

Let t_1 be the time required to go from the set (V_1) to (V_2).

All times must be positive.

Furthermore, let

$$z_{N-R} = \exp(t_{N-R}) \tag{36}$$

and since each time was assumed positive, note that each

$$z_j > 1; \quad \text{for } j = 1, 2, \cdots, N-1. \tag{37}$$

The variables (V_1) of the switching hypersurface satisfy the following equations:

$$v_{k,1} + \Delta^* = 2\Delta^* z_1{}^{sk} - 2\Delta^*(z_1z_2)^{sk} + 2\Delta^*(z_1z_2z_3)^{sk} - \cdots + 2\Delta^*(z_1z_2 \cdots z_{N-4})^{sk} - 2\Delta^*(z_1z_2 \cdots z_{N-3})^{sk} + 2\Delta^*(z_1z_2 \cdots z_{N-2})^{sk} - \Delta^*(z_1z_2 \cdots z_{N-1})^{sk}$$

$$\text{for } N \text{ odd}, k = 1, 2, \cdots, N \text{ and } z_j > 1. \tag{38}$$

The derivation of the above equation follows the same procedure as the derivation of (33-n). See Appendix II for the derivation.

Let

$$w_1 = z_1$$

$$w_2 = z_1z_2$$

$$w_3 = z_1z_2z_3$$

. . . . .

$$w_{N-1} = z_1z_2z_3 \cdots z_{N-2}z_{N-1}. \tag{39}$$

Substituting (39) into (38) and dividing by Δ^*, the equations that the variables (V_1) of the switching hyper-

surface must satisfy are: For N odd,

$$\frac{v_{k,1}}{\Delta^*} + 1 = 2w_1^{s_k} - 2w_2^{s_k} + 2w_3^{s_k} - \cdots - 2w_{N-3}^{s_k} + 2w_{N-2}^{s_k} - w_{N-1}^{s_k}$$

$$\text{for } \Delta^* = \pm 1,\ k = 1, 2, \cdots, N; \tag{40-a}$$

for N even,

$$\frac{v_{k,1}}{\Delta^*} - 1 = -2w_1^{s_k} + 2w_2^{s_k} - 2w_3^{s_k} + \cdots - 2w_{N-3}^{s_k} + 2w_{N-2}^{s_k} - w_{N-1}^{s_k}$$

$$\text{for } \Delta^* = \pm 1,\ k = 1, 2, \cdots, N. \tag{40-b}$$

Note that from the restriction of positive time and from the definition of the w_i, it follows that the w variables must satisfy the following restrictions:

$$w_{N-1} > w_{N-2} > \cdots w_3 > w_2 > w_1 > 1. \tag{41}$$

The restrictions imposed by (41) will be used to determine the correct Δ^*. These restrictions eliminate all possible trajectories in negative time from the mathematical expressions.

Eqs. (40-a) or (40-b) are the parametric equations of the switching hypersurface. The total number of these equations is N. The total number of the w variables is $N-1$. The restrictions (41) must be satisfied. Thus, given an initial point (V_0), then,

a) either the initial point is on the set (V_1), that is, the initial point lies on the switching hypersurface, or,
b) the initial point (V_0) does not belong to (V_1), that is, the point is not on the switching hypersurface.

Let

$$v_{1,0} = v_{1,1}, v_{2,0} = v_{2,1}, \cdots, v_{N-1,0} = v_{N-1,1}. \tag{42}$$

That is, use the first $N-1$ coordinates of the initial condition point, also called the *state* point, in the first $N-1$ equations given by (40-a) or (40-b).

For N odd, since $1/\Delta^* = \Delta^*$

$$\Delta^* v_{1,0} + 1 = 2w_1^{s_1} - 2w_2^{s_1} + \cdots + 2w_{N-2}^{s_1} - w_{N-1}^{s_1}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\Delta^* v_{N-1,0} + 1 = 2w_1^{s_{N-1}} - 2w_2^{s_{N-1}} + \cdots + 2w_{N-2}^{s_{N-1}} - w_{N-1}^{s_{N-1}}. \tag{43}$$

Since the coordinates of the state point (V_0) are known, one may solve (43) in the following manner:

1) Assume that $\Delta^* = +1$.
2) Solve for $w_1, w_2, \cdots, w_{N-1}$.
3) Check if the w variables determined satisfy the restrictions of (41).
4) If the restrictions of (41) are not satisfied, then use $\Delta^* = -1$ and repeat steps 1 and 2. A solution will always exist because of the property of the switching hypersurface's being continuous, infinite in extent, and separating the space into two parts.

Thus, from (43) and from the restrictions (41) one may solve for $w_1, \cdots, w_{N-1}$ and the correct Δ^*.

Since only $N-1$ equations out of N were used, then, one may use the w variables just determined and the correct Δ^* in the last equation

$$\Delta^* v_{N,1} + 1 = 2w_1^{s_N} - 2w_2^{s_N} + \cdots + 2w_{N-2}^{s_N} - w_{N-1}^{s_N} \tag{44}$$

and (44) yields the value of $v_{N,1}$.

Thus,

a) If (V_0) belongs to (V_1), then, $v_{N,1}$ determined from (44) is such that

$$v_{N,1} = v_{N,0}.$$

b) If (V_0) does not belong to (V_1), then,

$$v_{N,1} - v_{N,0} \neq 0. \tag{45}$$

The above procedure illustrates the methods that must be followed in order to test if the state point is on the switching surface.

The Concept of the Distance Function

The distance function is defined as the distance from the present state point to the switching hypersurface. This distance function has a magnitude and a polarity, which depends on whether the state point is "above" or "below" the switching hypersurface. The magnitude of the distance function may be the Euclidean distance from the state point to the hypersurface along some axis of the N-dimensional space or any linear or nonlinear single-valued function of it.

The distance function, as defined above, has the following properties:

a) If the state point is on the switching hypersurface, then the distance function is zero.
b) If the state point is not on the switching hypersurface, then the distance function will have either a positive or a negative polarity. Due to these properties, the distance function may serve directly as the output of the nonlinear computer, the manipulated variable $m(t)$ of Fig. 1.

Thus,

a) If (V_0) does not belong to (V_1), then $m(t)$ is the distance function from (V_0) to (V_1).
b) If (V_0) belongs to (V_1) but not to (V_2), then the distance function is the distance from (V_0) to (V_2).
c) If (V_0) belongs to (V_2), and, hence to (V_1), but not to (V_3), the distance function is the distance from (V_0) to (V_3), and so on.

The Generation of the Distance Function

The distance function, which is $m(t)$, the manipulated variable of Fig. 1, may be defined as follows:

Let

$$m(t) = -(-1)^N(v_{N,0} - v_{N,1}). \tag{46}$$

Note that the output of the relay $F = -\Delta = \text{sgn } m(t)$. The relation between $m(t)$ and the difference $v_{N,0} - v_{N,1}$ was obtained as follows: Given a point (V_0) such that $v_{N,0} - v_{N,1} > 0$, then one may find whether $\Delta = +1$ or $\Delta = -1$ will cause the trajectory to intersect the switching hypersurface in positive and finite time. The identical procedure was repeated for a point (V_0) such that $v_{N,0} - v_{N,1} < 0$ and it was found that (46) was the correct relationship.

Thus for systems described by an even-order differential equation, if the state point is above the switching hypersurface (above referring to a direction along the v_N axis), the output of the relay must be -1, while if the state point is below the switching hypersurface, the output of the relay must be $+1$.

If the system is described by an odd-order differential equation, then if the state point is above the switching hypersurface the output of the relay must be $+1$, and if the state point is below the switching hypersurface, the relay output must be -1.

Some Practical Considerations

In a real system, one must design a computer which will be able to solve (43) and take into account the restrictions given by (41). This solution of the algebraic equations will take some small but finite time. Moreover, since time derivatives of the error must be obtained for a high-order system, there exists noise and the values of the different variables are not known very accurately. Thus, the state point cannot be brought exactly on the switching surface but very close to it. This corresponds to the case where after the initial trajectory the following trajectories of the system will be close to the switching hypersurface. The distance function will have the correct polarity but its magnitude will be small. Fig. 6 shows, in a block-diagram form, the different computations required for the generation of the distance function $m(t)$.

The output is subtracted from the input, which was assumed to be zero, to form the error. The error is operated on by the ideal all-zero network and the individual lag networks to construct the state variables $v_1, v_2, \cdots, v_N$ as shown in Fig. 5.

The variables $v_1, \cdots, v_{N-1}$ are used as the known quantities in the solution of the nonlinear algebraic equations described by (43) to determine the auxiliary variables w which must satisfy the restrictions of (41). The variables w are used for the calculation of the switching surface variable $v_{N,1}$ according to (44). The difference between the computed variable $v_{N,1}$ and the

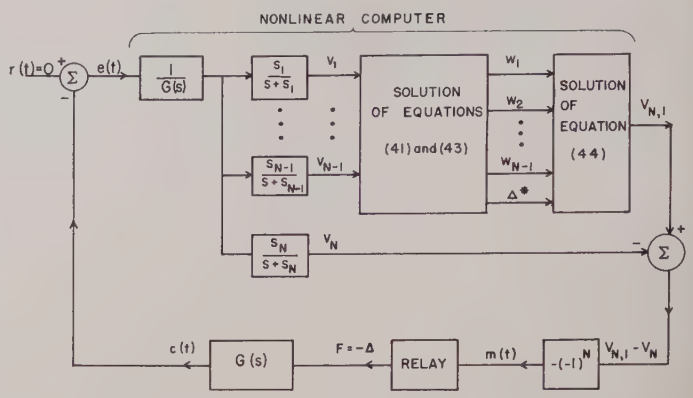


Fig. 6—Symbolic representation of suggested design procedure.

remaining state variable $v_{N,0}$ is used to compute the manipulated variable $m(t)$ as shown by (46). The variable $m(t)$ determines the present relay output.

If the order of the system is small, that is, for $N = 2$ or $N = 3$, it is possible to solve the nonlinear algebraic equations given by (43) explicitly and to determine a criterion based upon the present values of the state point which satisfies simultaneously the restrictions of (41) and yields the correct value of Δ^*. If this is the case, then the values of the w variables may be determined explicitly as functions of the state variable. One may then proceed to determine the distance function as an explicit function of the coordinates of the state point. If the system is of higher order and/or the poles of the system are not simple integers, then the solution of the nonlinear equations (43) and (41) must be performed by a computer in an "educated" trial-and-error fashion. It must be mentioned that due to the inequalities $w_{N-1} > w_{N-2} > \cdots > w_2 > w_1 > 1$ and $s_N > s_{N-1} > \cdots > s_2 > s_1 > 0$ certain terms in (43) are much larger than others and this information may be used in order to determine a rapidly convergent trial-and-error solution. Preliminary calculations on third- and fourth-order systems indicate that if the initial guess is $w_1 = 1$, then one may converge to the true solution with 3 to 6 tries, which may be performed quite rapidly in a medium-speed digital computer. See Appendix III for computations for a third-order system.

Conclusions

The equations for the switching hypersurface of the Nth-order maximum-effort bang-bang control system have been developed. The equations of the switching hypersurface were developed by considering the series of the optimal trajectories which bring the present state point to zero in the minimum possible time. The assumptions used were that the optimal solution existed and that the origin of the phase space could be reached with, at most, $N-1$ switchings (reversals) of the relay.

The equations that describe the relationship between the phase-space variables at the instant of the first switching of the relay were obtained in a parametric

form with a nonlinear function of time acting as the parameter of the equations. From the equations of the switching hypersurface, a convenient mathematical measure was obtained, the distance function, which described the position of the state point with respect to the switching hypersurface.

The main contribution of this paper is that the exact mathematical operations and computations that must be performed by the computer have been exactly specified. No memory about the previous behavior of the system was assumed and no feedback from the output of the relay was used in the mathematical computations. The latter is especially important in the stability of the system, because although the present output of the relay may be incorrect, the correct distance function of the system is applied to the relay such that the relay output is forced to its correct value. The concept of the switching curve for second-order systems has been used extensively before. The concept of the switching surface has been used[9,13,17,19] only for special cases. The scope of this paper was to extend the available theory to an arbitrary-order linear system and to use the integrated concept of the switching hypersurface and of the distance function for the actual design of a time optimal control system.

The theory and method of design presented may be extended to the case in which the terminal point is not the origin, but is any other arbitrary point in the phase space.[20] Only the equation of the final trajectory need be modified to include the coordinates of the terminal point, provided the system is controllable.

If the input to the system is not zero, for example, the input may be composed of steps, ramps, and other polynomial-type terms; then, the equation of the switching hypersurface will depend on the present and predicted value of the input function and of its derivatives. Identical linear transformations may be performed on the error variables; however, the nonlinear algebraic equations which describe the switching hypersurface will include not only the coordinates of the state point but also some transformed variables of the input function.[20]

Furthermore, in the case of inputs, the system itself must be controllable, that is, the inputs must be limited to those that the system can follow without any steady-state error in a finite transient time.[20]

Appendix I

Derivation of the v Variables

Consider (22)

$$\dot{x}_j(t) = -s_j x_j + b_{jN}\Delta. \tag{47}$$

Take Laplace transforms

$$sx_j(s) = -s_j x_j(s) + b_{jN}\Delta(s) \tag{48}$$

$$(s + s_j)x_j(s) = b_{jN}\Delta(s)$$

$$\therefore \quad x_j(s) = \frac{b_{jN}\Delta(s)}{(s + s_j)}. \tag{49}$$

But from (24)

$$v_j(s) = x_j(s)\frac{s_j}{b_{jN}}$$

$$\therefore\ v_j(s) = \frac{s_j}{(s + s_j)}\Delta(s). \tag{50}$$

But from (5)

$$\Delta(s) = -\frac{C(s)}{G(s)}$$

$$\therefore\ v_j(s) = -\frac{1}{G(s)}\frac{s_j}{(s + s_j)}C(s). \tag{51}$$

This operation is shown in Fig. 5.

Appendix II

Derivation of the Equations of the Switching Hypersurface

The set (V_{N-1}) is described by the equation

$$0 = (v_{k,N-1} - \Delta^*)e^{-s_k t_{N-1}} + \Delta^*$$

or

$$v_{k,N-1} = -\Delta^* e^{s_k t_{N-1}} + \Delta^* \quad k = 1, \cdots, N. \tag{52}_{N-1}$$

The set (V_{N-2}) is described by the equation

$$v_{k,N-1} = (v_{k,N-2} + \Delta^*)e^{-s_k t_{N-2}} - \Delta^*$$

or

$$v_{k,N-2} = (v_{k,N-1} + \Delta^*)e^{s_k t_{N-2}} - \Delta^* \tag{52}_{N-2}$$

. . . . . .

The set (V_1) is described by the equation

$$v_{k,2} = (v_{k,1} + \Delta^*)e^{-s_k t_1} - \Delta^*$$

or

$$v_{k,1} = (v_{k,2} + \Delta^*)e^{s_k t_1} - \Delta^*. \tag{52}_1$$

Using (36)

$$v_{k,N-1} = -\Delta^* z_{N-1}^{s_k} + \Delta^* \tag{53}_{N-1}$$

$$v_{k,N-2} = (v_{k,N-1} + \Delta^*)z_{N-2}^{s_k} - \Delta^* \tag{53}_{N-2}$$

. . . . . .

$$v_{k,1} = (v_{k,2} + \Delta^*)z_1^{s_k} - \Delta^* \quad N \text{ odd } k = 1, \cdots, N. \tag{53}_1$$

By successive elimination of the $v_{k,N-1}, v_{k,N-2}, \cdots$ variables, equation (38) is obtained.

[19] O. J. M. Smith, "Feedback Control Systems," McGraw-Hill Book Co., Inc., New York, N.Y.; 1958.

[20] M. Athanassiades, "Theory and Design of Bang-Bang Control Systems," Ph.D. dissertation, University of California, Berkeley; June, 1961.

Appendix III

Experimental Results for a Third-Order System

Let $N=3$. Then the switching surface satisfies the equations

$$\Delta^* v_{1,1} + 1 = 2w_1^{s1} - w_2^{s1} \tag{54}$$

$$\Delta^* v_{2,1} + 1 = 2w_1^{s2} - w_2^{s2} \tag{55}$$

$$\Delta^* v_{3,1} + 1 = 2w_1^{s3} - w_2^{s3} \tag{56}$$

and

$$w_2 > w_1 > 1; \qquad s_3 > s_2 > s_1 > 0. \tag{57}$$

Given $v_{1,1}$ and $v_{2,1}$ one may solve for $v_{3,1}$. Therefore, one needs to solve only (54) and (55).

If w_1 is chosen then one may solve for w_2 from (54). Define an error

$$E = 2w_1^{s2} - w_2^{s2} - (\Delta^* v_{2,1} + 1). \tag{58}$$

If $E=0$, then the w_1 chosen is the correct one. From (58)

$$\frac{\partial E}{\partial w_1} = 2s_2 w_1^{s2-1} - s_2 w_2^{s2-1} \frac{\partial w_2}{\partial w_1}. \tag{59}$$

But from (54)

$$s_1 w_2^{s1-1} \frac{\partial w_2}{\partial w_1} = 2s_1 w_1^{s1-1}. \tag{60}$$

Substituting in (59)

$$\frac{\partial E}{\partial w_1} = 2s_2 w_1^{s2-1} - 2s_2 w_2^{s2-1}(w_1^{s1-1}/w_2^{s1-1}) \tag{61}$$

If $w_2 > w_1$ and $s_2 > s_1$ then

$$\frac{\partial E}{\partial w_1} < 0 \quad \text{independent of } \Delta^*. \tag{62}$$

Based on the above result it was found that the following computational procedure converges to the true value of w_1 and w_2.

1) Assume $\Delta^* = \operatorname{sgn} v_{1,1}$.
2) As an initial choice for w_1, say w_{10}, choose $w_{10} = w_{20}$ such that $w_{10}^{s1} = (\Delta^* v_{1,1} + 1)$.
3) Solve for w_{10} and $w_{20} = w_{10}$.
4) Substitute w_{10} and w_{20} in (58) and evaluate E_0, where $E_0 = 2w_{10}^{s2} - w_{20}^{s2} - (\Delta^* v_{2,1} + 1)$.
5) If E_0 is negative then the choice of Δ^* is wrong. If E_0 is positive, the choice of Δ^* is right. If E_0 is negative then an increase in the value of w_{10} will result in larger negative error and the process would diverge.
6) If the error E_0 is negative, change the value of Δ^* found in step 1) by its negative. Then assume that $w_{10} = 1$.
7) Change w_{10} by an increment based in some estimate of the slope of the E vs w_1 curve near zero error.

It was found that the above iterative process is efficient and that the number of iterations required to determine w_1 and w_2, with good accuracy, ranged between 3 and 7.

Three sample calculations are shown below:

Example 1

$s_1 = 1.2$ $s_2 = 1.8$
$v_{1,1} = 2.0$ $v_{2,1} = -5.0$

	w_1	w_2	E	Δ^*
1st iteration	2.4981	2.4981	9.1962	1.0
2nd iteration	4.6051	6.5281	5.9686	
3rd iteration	5.6962	8.5503	2.2261	
4th iteration	6.0679	9.2349	0.6721	
5th iteration	6.1770	9.4355	0.1896	
6th iteration	6.2076	9.4916	0.0527	

Example 2

$s_1 = 1.2$ $s_2 = 1.8$
$v_{1,1} = 2.0$ $v_{2,1} = -20.0$

	w_1	w_2	E	Δ^*
1st iteration	2.4981	2.4981	24.1960	1.0
2nd iteration	8.0421	12.8460	5.2207	
3rd iteration	8.7846	14.1960	0.3843	
4th iteration	8.8369	14.2910	0.0264	

Example 3

$s_1 = 1.2$ $s_2 = 3.2$
$v_{1,1} = -5$ $v_{2,1} = -5$

	w_1	w_2	E	Δ^*
1st iteration	4.4510	4.4510	112.8700	−1.0
2nd iteration	5.8566	7.1970	12.9221	
3rd iteration	5.9324	7.3425	0.3845	
4th iteration	5.9346	7.3466	0.0127	

A third-order system with $s_1=1.2$, $s_2=1.8$, $s_3=2.2$, was simulated on a Bendix G-15 computer. The experimental results were in accordance with the theory.

Model Feedback Applied to Flexible Booster Control*

G. E. TUTT† AND W. K. WAYMEYER†, MEMBER, IRE

Summary—The problem of the filtering of elastic phenomena (bending and sloshing) from space booster-control feedbacks has become increasingly difficult as bending frequencies approach control frequencies in large vehicles. The design problem is compounded by the loop-gain requirements for control of an aerodynamically unstable airframe in a dynamic wind environment.

This paper presents a "model feedback" approach to this design problem in which the system does not attempt to "adapt" to body bending, but instead is contrived to ignore it. The vehicle-body feedbacks, required for stability and control, are synthesized by a combination of actual body motion and information from a model of the rigid body acted upon by the control force.

Following an investigation of stability and response for this system, attention is focused on the problem of response to disturbances such as wind, gusts and shear.

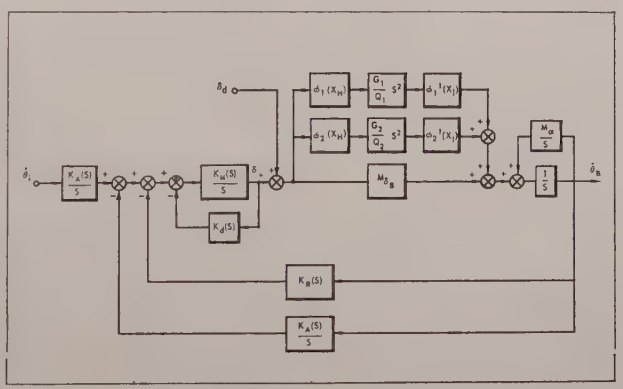
Introduction

THE fundamental control problems in the design of flexible space-vehicle boosters center around the control of an aerodynamically unstable airframe in a dynamic wind-shear (jet stream) environment. The control-loop gains required to prevent tumbling are usually high enough to make the filtering of bending and sloshing feedbacks difficult. As the structural efficiency of the tankage is improved and the body-bending frequencies approach the controlled airframe frequency, new approaches to control design are needed. A common suggestion is that "adaptive" control systems should be developed to cope with this problem. In this paper, a feedback-model approach to this design problem, which shows promise, is outlined in some detail. This approach does not adapt to body bending, but instead is contrived to ignore it.

The designer's knowledge of thrust, vehicle weight, moment of inertia, and aerodynamic characteristics is usually quite good. The approach developed here does not preclude, however, a desire to know them better, or to "adapt" to them as the booster rises through the sensible atmosphere; rather this approach is concerned with the control of the nonrigid-body aspects of the vehicle. For the case in which the rigid body parameters are not well known, adaptive techniques might be applied for their identification and incorporation.

Conventional System

The conventional method of booster control is a proportional attitude control system with rate feedback for rigid-body stability, and control force derived from gimballing a thrust vector. The problem then is to design compensation and set gains, such that nowhere in flight will a rigid body or resonance instability occur. Fig. 1 shows this type of system as used in a typical



$K_A = 1.37 \quad K_R = 0.563$

$$\frac{G_1}{Q_1} = \frac{48.9}{S^2 + 0.343S + 600}$$

$K_M = 14.7 \quad M_{\delta B} = 4.76 (t = 0)$
$= 5.53 (t = 60 \text{ seconds})$

$K_d = 0.85 \quad M_\alpha = 0 \ (t = 0)$
$= 3.06 \ (t = 60 \text{ seconds})$

$$\frac{G_2}{Q_2} = \frac{97.9}{S^2 + 0.669S + 1120}$$

Fig. 1—Conventional autopilot showing some typical gains.

* Received by the PGAC, December 15, 1960; revised manuscript received, March 15, 1961.

† Douglas Aircraft Co., Inc., Santa Monica, Calif.

Thor space booster. The block diagram shows two modes of body-bending and aerodynamic effects, simplified to ignore lateral motion components. The vehicle dynamics are thus simplified to include only one important nonrigid-body effect (uncoupled bending of the airframe), so as to avoid complication which would not be pertinent to a demonstration of the basic model-feedback principle. Although propellant sloshing is a very important effect, there exist methods of stabilization (baffles) which do not require major changes in the airframe itself. Consequently, as presented here, the emphasis is placed upon control of the vehicle body bending. The feedback blocks are shown as functions of S to indicate the existence of some form of compensation. Command and disturbance transfer functions are, respectively:

$$\frac{\theta_B}{\theta_i} = \frac{\dfrac{K_A}{S}\,\dfrac{K_M}{S}\,\dfrac{M_{\delta B}(S)}{S}}{1 + \dfrac{K_M(S)}{S}\left(K_D(S) + \dfrac{M_{\delta B}(S)}{S}K_R(S) + \dfrac{M_{\delta B}(S)}{S}\,\dfrac{K_A(S)}{S}\right)}$$

and

$$\frac{\dot\theta_B}{\delta_d} = \frac{\dfrac{M_{\delta B}(S)}{(S)}}{1 + \dfrac{M_{\delta B}(S)}{S}\left(\dfrac{K_M(S)}{1 + \dfrac{K_M K_D(S)}{S}}\right)\left(K_R(S) + \dfrac{K_A(S)}{S}\right)}.$$

where

$$M_{\delta B}(S) = \left(M_{\delta B} + \phi_1(X_H)\,\frac{G_1}{Q_1}S^2\phi_1'(X_I) + \phi_2(X_H)\,\frac{G_2}{Q_2}S^2\phi_2'(X_I)\right)\left(\frac{S}{S^2 - M_\alpha}\right).$$

Clearly the amplitude and phase of the bending feedback will depend upon the slope of the mode (n) at the instrument location $\phi_n'(X_I)$, and the characteristics of the compensation as a filter at the modal frequency. The filters cannot efficiently reject the lower bending-mode frequencies if these are close to control frequencies, since a rejection of control frequencies would impair system response.

Model Feedback

The idea of a model feedback system is to allow heavy filtering of the bending frequencies, while the rejected rigid-body information is returned to the feedback via a model of the rigid body only. Of course, very low-frequency elastic-body information is retained to maintain long-term attitude accuracy. This basic configuration is shown in Fig. 2, where it can be seen that the

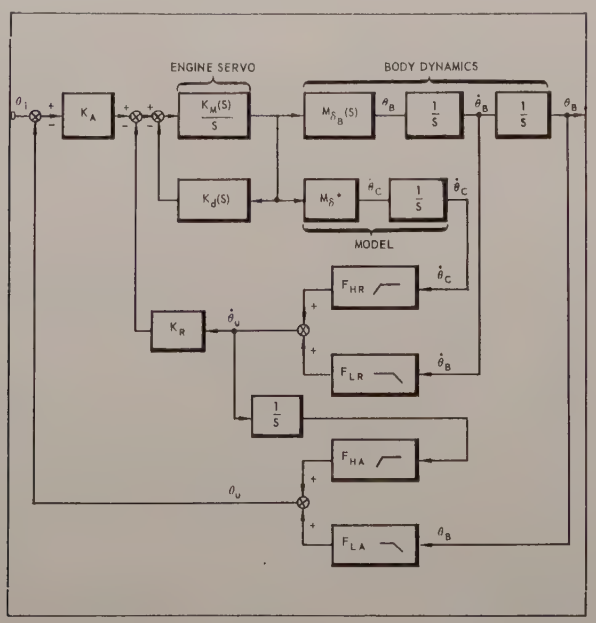


Fig. 2—Basic model feedback system.

engine position δ operates upon the vehicle-control moment effectiveness $M_{\delta_B}(s)$ and upon a model of the vehicle as a rigid body, $M_\delta{}^*$. Thus the total rigid-body feedback would be approximately as before, with a high rejection of body bending.

This approach depends upon a knowledge of the rigid-body control effectiveness (thrust, moment of inertia, c.g. location, etc.), all of which are usually predictable for a space booster. The assumption made for the basic system of Fig. 2 is that all vehicle angular motion arises from thrust vector deflection as measured at the engine position instrument. This is strictly true, of course, only for the nondisturbance, zero q (dynamic pressure) condition.

Fig. 3 shows the MF (model feedback) diagram partitioned as for the conventional system, with the addition of a model of the aerodynamic moment effectiveness.

The command and disturbance transfer functions are:

$$\frac{\theta_B}{\theta_i} = \frac{\dfrac{K_A}{S}\dfrac{K_M}{S}\left(\dfrac{M_{\delta B}(S)S}{S^2 - M_\alpha}\right)}{1 + \dfrac{K_M K_d}{S} + \dfrac{K_M M_\delta{}^*}{S}\left[\dfrac{K_R(S+\omega) + K_A}{(S+\omega)^2 - M_\alpha{}^*}\right] + \dfrac{K_M M_\delta(S)\omega}{S(S^2 - M_\alpha)}\left[\dfrac{K_R S(S+\omega) + K_R M_\alpha{}^* + K_A(2S+\omega)}{(S+\omega)^2 - M_\alpha{}^*}\right]}$$

$$\frac{\dot\theta_B}{\delta_d} = \frac{\dfrac{M_{\delta B}(S)S}{S^2 - M_\alpha} + \dfrac{M_{\delta B}(S)K_M}{S^2 - M_\alpha}\left\{K_d + M_\delta{}^*\left[\dfrac{K_R(S+\omega) + K_A}{(S+\omega)^2 - M_\alpha{}^*}\right]\right\}}{\text{Same Denominator as Above}},$$

where $M_{\delta_B}(S)$ includes the bending terms.

Comparing this command transfer function with that for the conventional system suggests that the MF block diagram could be rearranged to appear like that for the conventional system. If this were done, the engine feedback would be transformed to

$$K_d + M_\delta{}^*\left[\frac{K_R(S+\omega) + K_A}{(S+\omega)^2 - M_\alpha{}^*}\right],$$

while the rate and attitude feedbacks would have multipliers involving only the filter frequency (ω) and the model parameter $M_\alpha{}^*$.

It may be seen from the transformed engine-feedback function that the inclusion of $M_\alpha{}^*$ in the model system is required for aerodynamic stability at any appreciable levels of M_α. The reason for this is that deletion of the $M_\alpha{}^*$ term would cause the model to revert to $M_\delta{}^*$, in which case the dc gain of the transformed engine feedback would become

$$K_d + M_\delta{}^*\left(\frac{K_R}{\omega} + \frac{K_A}{\omega^2}\right).$$

Quantitatively, if ω were 2 radians per second, for example, using the gains from Fig. 1, this function would then be about four times greater than K_d alone. The resulting drop in dc control gain (δ/θ) would produce a closed-loop pole in the right-half S plane when

$$4M_\alpha \gtrsim M_\delta \frac{K_A}{K_d}.$$

However, it is not possible for $M_\alpha{}^*$ to operate upon only computed information, since a closed-loop pole-zero pair would appear in the right-half plane and these would cancel only if $M_\alpha{}^* = M_\alpha$ exactly. Hence, one gets the configuration shown in Fig. 3, where $M_\alpha{}^*$ is acted upon by the combination of computed and actual body position (using the simplifying assumption that $\theta = \alpha$).

Comparison with Conventional Sytem

A direct comparison, in terms of stability and response, will now be made between the conventional system of Fig. 1, and the MF system of Fig. 3. The same

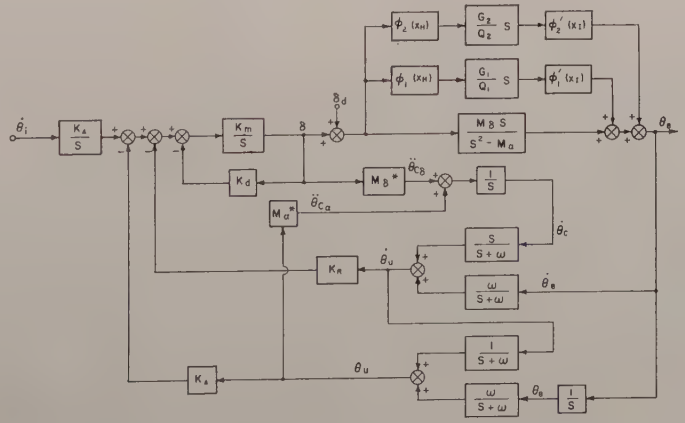


Fig. 3—Model feedback with two bending modes.

gains and bending dynamics were used for both, with the first- and second-mode bending frequencies at 3.9 and 5.3 cycles per second, respectively.

Nyquist diagrams of the conventional θ_B/θ_i characteristic function are plotted in Figs. 4 and 5 for two flight conditions. For Fig. 4 the rate and attitude gyros were well forward of the first mode antinode, while for Fig. 5 they were well aft. These conditions might correspond to times early in flight and late in flight, respectively. The corresponding Nyquists for the MF system are Figs. 6 and 7. It can be seen that while a bending instability occurred in the conventional system when the bending-mode shapes changed with respect to the instrument locations, no such problem was encountered in the MF system. As expected, bending rejection was high in the MF configuration.

The response to commands (in the form of step rates) is shown for both the conventional and the MF systems for a liftoff condition (no aerodynamic forces) in Fig. 8, and for a region of flight near the maximum q in Fig. 9. These plots show that response has not deteriorated with respect to the conventional system, which is not surprising, since rigid-body feedback is essentially the same when $M_\delta = M_\delta{}^*$ and $M_\alpha = M_\alpha{}^*$.

A comparison of disturbance response is made by introducing a step 2660-pound force at the gimbal location for each system. The comparative responses are shown in Fig. 10 for $M_\alpha = 0$ (no aerodynamics) and in Fig. 11 for $M_\alpha = 3.06$ (maximum q).

A very poor response to disturbances is thus evident for the MF system, as was to be expected, since body motion is not generated solely by engine position. In this case, the engine responds only to information passed through the heavily-filtered real body loop.

For the plots of MF response at $M_\alpha = 3.06$, a mapping of the closed-loop roots is shown in the lower right corner. In the case of the command transfer function, the pole at 0.25 is cancelled by a zero if $M_\alpha = M_\alpha{}^*$, leaving the dominant roots to be the conjugates shown. The effect of this "high time-constant" pole depends then upon how close $M_\alpha{}^*$ is to M_α. No such cancellation exists for the disturbance transfer function, however (see Fig. 11), and the pole at 0.25 becomes a dominant root, hence the slow response. Figs. 12 and 13 show command re-

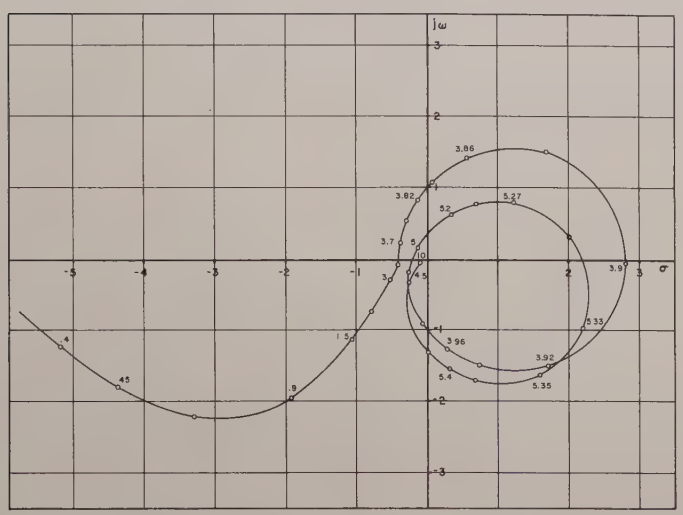


Fig. 4—Nyquist plot of conventional system—first flight condition.

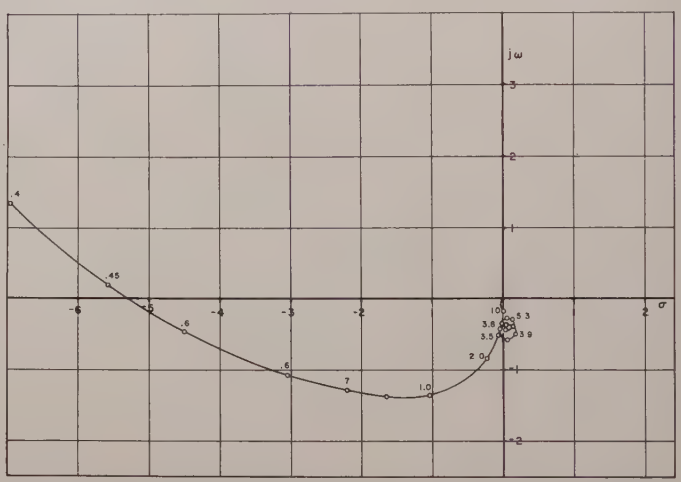


Fig. 6—Nyquist plot of model feedback system—first flight condition.

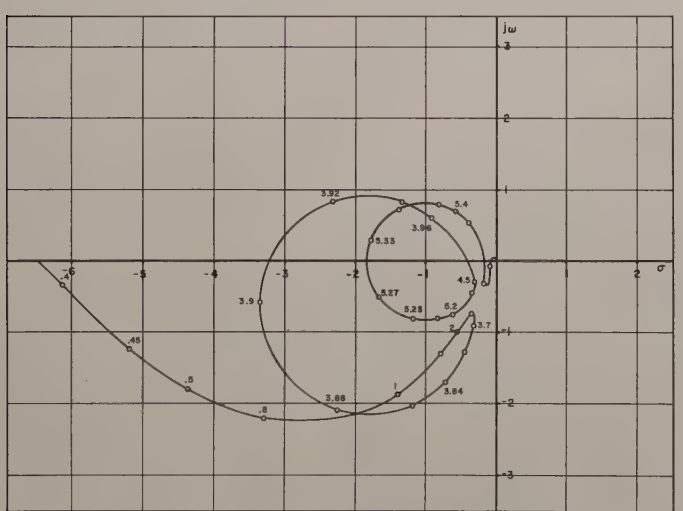


Fig. 5—Nyquist plot of conventional system—second flight condition.

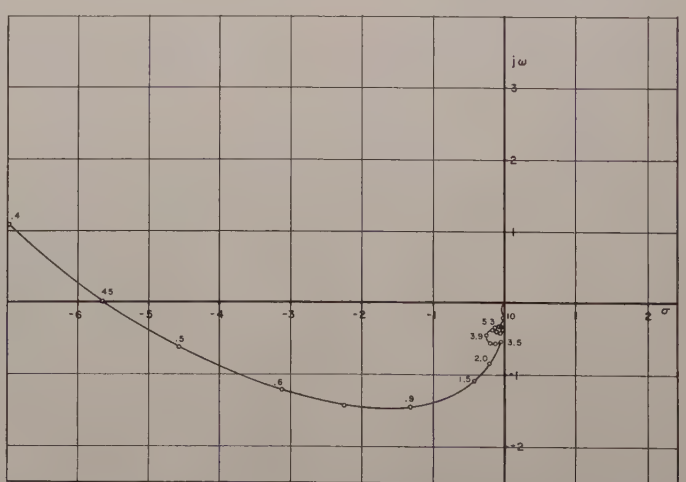


Fig. 7—Nyquist plot of model feedback—second flight condition.

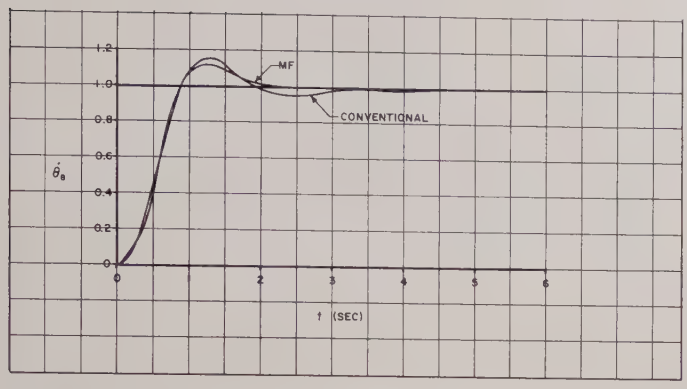


Fig. 8—Rate response to step-rate command—liftoff condition.

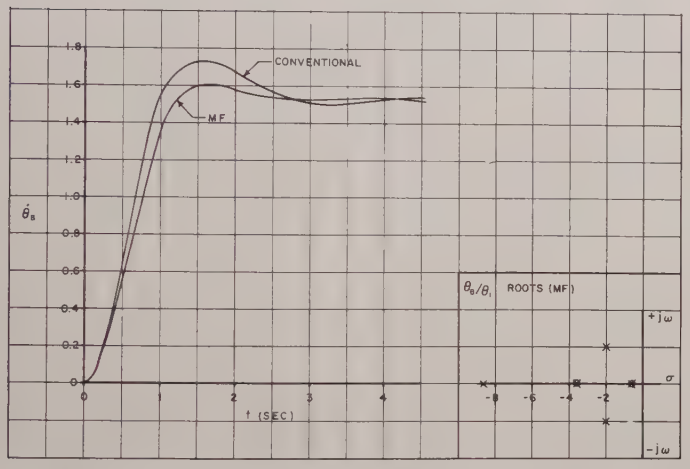


Fig. 9—Rate response to step-rate command—maximum q condition.

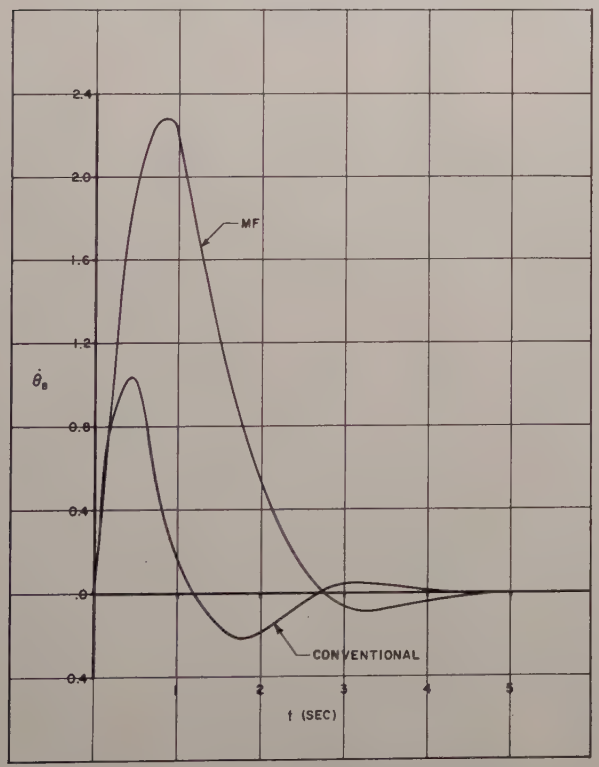


Fig. 10—Rate response to step-disturbance force (2660@X_H) for the liftoff condition.

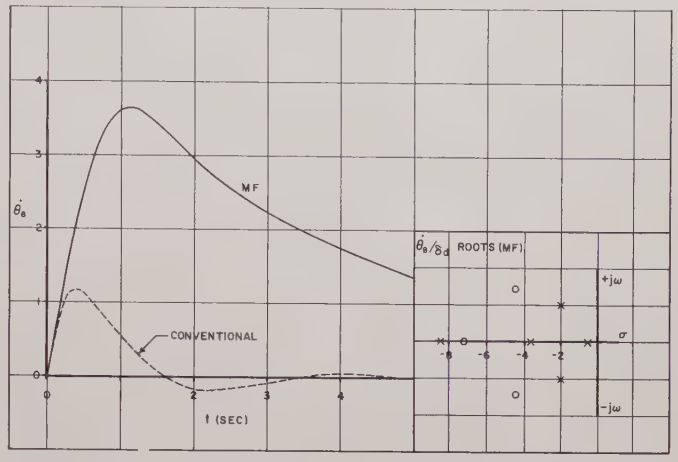


Fig. 11—Rate response to step-disturbance force—maximum q condition.

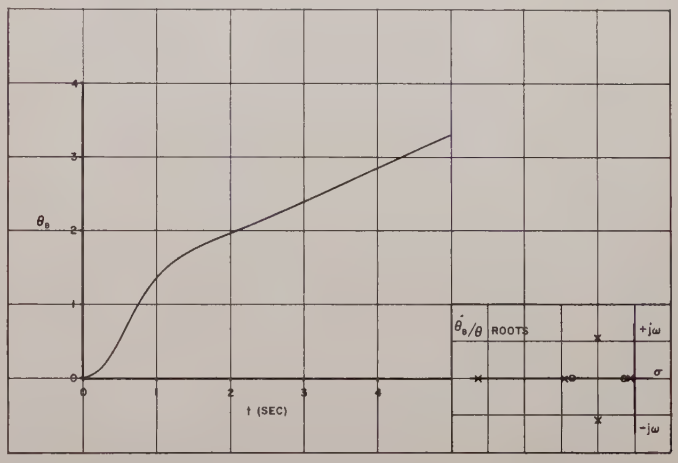


Fig. 12—MF command response with $M_\alpha{}^* = 15$ per cent less than M_α.

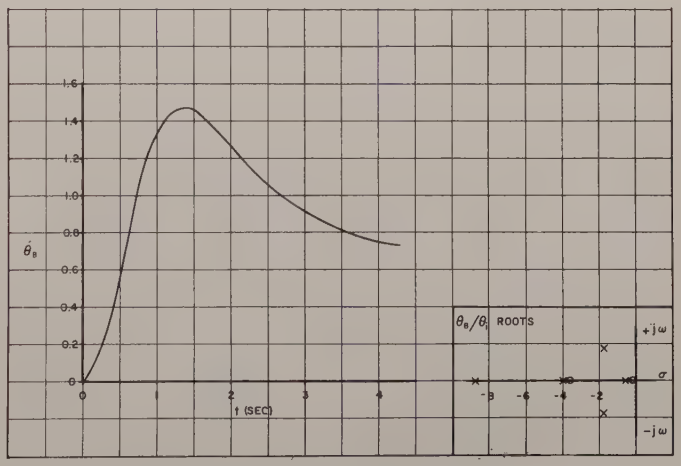


Fig. 13—MF command response with $M_\alpha{}^* = 15$ per cent greater than M_α.

sponse for $M_\alpha{}^*$ 15 per cent less than M_α, and 15 per cent greater, respectively. These plots show that tolerances placed on $M_\alpha{}^*$ (with respect to M_α) should be heavily weighted toward the positive ($M_\alpha{}^* > M_\alpha$). Indeed it might be well never to allow $M_\alpha{}^*$ to become less than M_α.

Modifications for Disturbance Response

Clearly some sort of modification to the MF system is required if anything but very low-frequency disturbances are to be tolerated. If no disturbances were anticipated, or if their fundamental harmonic component were below the filter-break frequency, then of course the system is adequate as it stands, and a logical simplification would probably be removal of the rate gyro.

Since, for booster flight through the atmosphere, the origin of disturbances is known to be aerodynamic, two additions to the basic MF system based on α-meter feedback will be mentioned. The problem is to obtain some sort of disturbance information without re-introducing the bending which was previously rejected. The first method involves the measurement of the angle of attack by a suitable α meter, and a combination of this with an approximate α as generated by the model loop. This is depicted in Fig. 14, which shows an expansion of the basic model to include lateral-motion terms which effect α. The idea here is that some portion of the initial disturbance information will be affected only by the filter, rather than the added integrations of body motion, and hence, better feedback signals of the disturbance will be provided.

To observe the response of this configuration, a step-wind velocity was applied (for the case of $M_\alpha = 3.06$) such that the disturbance moment on the vehicle was the same as for the step 2660-pound force at the gimbal location. This velocity (58.7 feet per second) produces an initial α of 1.81°, hence initially

$$\ddot{\theta} = (1.81°) \times \left(3.06 \frac{\text{deg/sec}^2}{\text{deg } \alpha}\right) = 5.53 \text{ deg/sec}^2.$$

This is the same as the previously applied moment, which was equivalent to 1° of engine deflection (2660 pounds at X_H). Fig. 15 shows that a significant improvement in disturbance response was obtained by addition of this low-frequency angle-of-attack information in the feedback. No significant change in bending relative stability would occur from this, since K_α was chosen such that $K_A' = K_\alpha + K_A$, where K_A' is the attitude loop gain previously used.

Thus it is demonstrated that a significant improvement in control-system response to disturbances (wind) is obtained when the angle of attack (in addition to inertial attitude) is fed back. This improvement, though significant, may not be good enough, as high-frequency information is filtered from the angle-of-attack data in order to suppress the sensing of structural vibration.

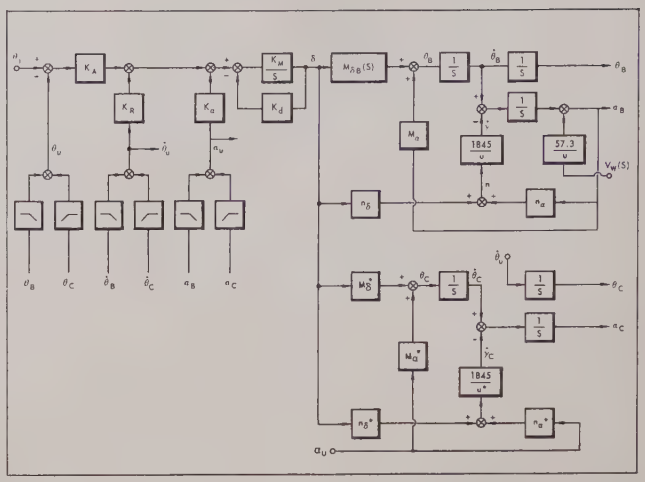


Fig. 14—Attitude angle-of-attack model feedback system.

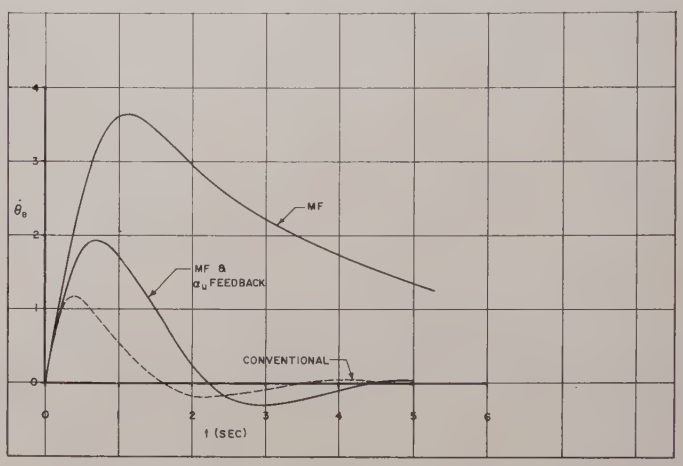


Fig. 15—Disturbance response showing the effect of α_u feedback.

A Second Method

It is therefore clear that a method of sensing the angle of attack (or the equivalent) without sensing bending would be of real value, since it would permit immediate recognition and feedback of wind inputs.

To explore how this might be done, the content of the acceleration, attitude rate and angle-of-attack transducer measurements is investigated for instruments mounted at the nose of the vehicle (Fig. 16).

The accelerometer measures the lateral acceleration of the center of gravity caused by aerodynamic forces ($N_\alpha\alpha$) and by the deflected thrust ($N_\delta\delta$). In addition, it measures angular acceleration of the vehicle in proportion to its distance from the c.g. ($1/1845\ddot{\theta}$) and oscillatory acceleration due to the bending of the airframe $\phi_1/32.2\ddot{X}_B$. The angle-of-attack meter measures the direction of motion of the vehicle nose with respect to the local wind

$$\left(\alpha_{\text{c.g.}} + \frac{1}{57.3U}\dot{\theta}\right)$$

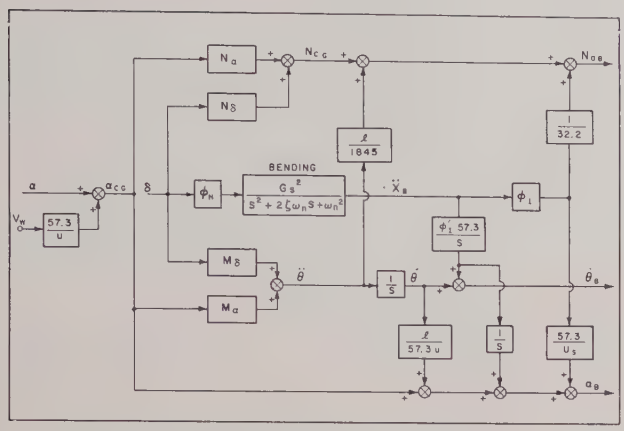


Fig. 16—Simplified dynamics to obtain $\dot{\theta}$, α and N_a of the vehicle as measured at the nose.

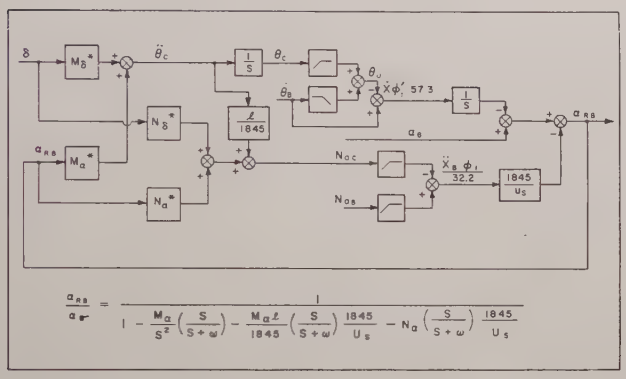


Fig. 17—A method of obtaining α (free of bending) to include wind inputs.

and is affected by body bending in two ways. Lateral velocity of the nose due to bending is interpreted as a change in local wind direction ($57.3\phi'/U\dot{X}_B$) and angular displacement of the vehicle nose because of bending ($X_B\phi_1'$ 57.3) is also sensed as an apparent rotation of the wind velocity. A device from which attitude rate at the nose is derived will sense bending ($\phi_1'57.3\dot{X}_B$) also.

It is true that the bending modes and frequencies are not always well known, and that the model feedback control presented in this paper is intended to eliminate any requirement for precise information in this area. It is, therefore, important to note that the bending sensed by the accelerometer is common to one component of the bending sensed by the angle-of-attack meter, except by a known constant and missile velocity. Similarly, the bending sensed by both the attitude-rate device and the angle-of-attack meter is the same.

These facts then suggest that bending information extracted from the accelerometers and rate instruments can be used to cancel bending information in the angle-of-attack data. Fig. 17 represents in block diagram form the concepts involved in this operation. Use is made of techniques presented in earlier sections. In essence, the technique is this: 1) Determine rigid-body rate and acceleration at the nose by a combination of techniques presented earlier, 2) subtract measured quantities from synthetic quantities, 3) subtract the difference from measured angle of attack, obtaining bending-free angle-of-attack information.

If the resulting α information is then used as an improved angle-of-attack feedback (as compared with that of Fig. 14), it is expected that the desired response to all forms of disturbance would be obtained.

Conclusions

To apply the concepts presented here to the control of a particular vehicle, certain quantitative aspects would require investigation. These include: 1) the determination of which model parameters are actually required in the model, 2) the sensitivity of the model parameter values to system performance, 3) determination of the most practical method of keeping the model parameter values within the required tolerances with respect to corresponding vehicle parameters, 4) for a particular vehicle and mission, the determination of the simplest and most practical form of α feedback for adequate response to the disturbances to be encountered, and 5) selection of a suitable filter frequency (ω) to give the required rejection of body bending.

A system has been derived, then, which virtually eliminates the possibility of a bending-mode instability even for wide variations in modal parameters, while maintaining the desired response to commands. This has been done at the expense of adding some complication to the system, in the form of model and α-feedback loops. It is also required that certain vehicle rigid-body dynamic parameters be identified, either during or previous to flight. However, in the design of a conventional control system, these dynamic parameters must be determined as well as the less readily obtainable bending parameters. Also, the compensation networks for the conventional systems could well be more complex, componentwise, than the MF model configurations.

Symbols and Nomenclature

F_H = High-pass filter
F_L = Low-pass filter
K_A = Attitude feedback gain
K_R = Rate feedback gain
K_M = Engine servo forward gain
K_d = Engine servo feedback gain
M_δ = Control moment effectiveness
M_α = Aerodynamic moment effectiveness
N_δ = Control lateral acceleration effectiveness
N_α = Aerodynamic lateral acceleration effectiveness
U = Vehicle velocity along the flight path
N_a = Lateral acceleration
c.g. = Center of gravity

l = Moment arm length
K_α = Angle-of-attack feedback gain
S = Laplace variable
$*$ = (Superscript) denotes a rigid-body model parameter
I = (Subscript) denotes a variable as measured at the instrument location
c = (Subscript) denotes a computed variable
u = (Subscript) denotes variable obtained from combination of actual and computed variables
B = (Subscript) denotes actual body variables or parameter
d = (Subscript) denotes a disturbance variable
θ = Body-attitude angle with respect to the gyro reference
δ = Engine deflection angle with respect to the body centerline
α = Aerodynamic angle of attack
γ = Flight path angle
$\phi_n(X_m)$ = Relative deflection of bending-mode n at location m
$\phi_n'(X_m)$ = Relative slope
ω_n = Natural frequency of mode n
ζ = Damping ratio
X_m = Deflection of station m in bending with respect to the rigid-body centerline
V_w = Lateral component of wind velocity.

Terminal Control System Applications*

E. A. O'HERN† AND R. K. SMYTH†, MEMBER, IRE

Summary—**This paper describes a synthesis procedure for terminal control systems. The synthesis procedure is illustrated by application to the design of an aircraft automatic landing system.**

A simplified model for the landing aircraft is chosen and justified. The prediction equations for this model are discussed.

Using the prediction equations, a two-condition terminal controller, which controls the altitude and altitude rate at time of touchdown to the desired values, is developed.

The terminal control landing system developed is simplified by combining feedback loops in a straightforward fashion to arrive at a practical mechanization. A stability analysis of the simplified terminal controller, considered as a servo with time variable gains and time constants, is presented.

The results of extensive analog computer simulation and flight test of the terminal control landing system developed in this paper are reviewed. It is shown that the synthesis procedures developed in this paper result in a system which can be implemented by state-of-art hardware to land aircraft successfully.

I. Introduction

A TERMINAL control system is one which controls the system state variables to specified values at some fixed, terminal time in the future. The concept of terminal control is based on the simple fact that if the initial conditions of a system with known differential equations of motion are specified, then the future response of the system can be predicted in the absence of disturbances to the system.

The terminal controller can be so designed that the system approaches the final state from the initial state in a desirable fashion. The desired final state is attained by exercising continuous control and continuously predicting the terminal conditions. Therefore, the system is controlled to the desired final value of the output variables even though disturbances occur. Appropriate control action is initiated to ensure at all times that the predicted terminal conditions match the specified terminal conditions.

The principal application of terminal control systems has been to automatic landing flare systems for aircraft. A number of system approaches have been employed for automatic landing of aircraft. These various approaches fall generally within three categories:

1) Altitude Rate Commanded as a Function of Altitude.
   The exponential flare, wherein altitude rate is commanded proportional to existing altitude, is the simplest and most prominent system of this category, in which longitudinal range of the aircraft path is ignored.
2) Fixed Path.
   Here the aircraft is controlled to follow a preset altitude-range trajectory.
3) Terminal Controller.
   Here the values of sinking rate and altitude at a particular time (corresponding to termination of flare) are predicted and controlled to acceptable values. This concept has the advantage of effecting some control of range by virtue of direct control of total flare time, without the attendant disadvantage of extending the control effort necessary to follow a preset trajectory throughout the

* Received by the PGAC, December 15, 1960; revised manuscript received, March 15, 1961.
† Armament and Flight Control Div., Autonetics, Downey, Calif.

flare. As a result of this requirement for less control action to achieve the desired results, it follows that recovery from disturbances is enhanced, thus minimizing resulting dispersions in terminal conditions.

It is the authors' experience that, of these various landing system approaches, the terminal controller landing system exhibits the greatest tolerance to disturbances and other off-nominal conditions. The net result is that the dispersion in the touchdown point and the sinking rate at touchdown is less than for the competing system approaches. It is also the experience of the authors that the terminal control landing system is considerably easier to fly manually than the competing systems when the pilot operates as a servo to maintain the terminal control system error zero.

Early investigations into terminal control concepts were initiated in early 1955 at the Massachusetts Institute of Technology Dynamic Analysis and Control Laboratory.[1–3] A large number of publications were made in which the prediction theory was developed and techniques for exercising terminal control were synthesized and analyzed. Shortly after this beginning, the authors and their colleagues at Autonetics initiated a program organized to develop further the basic concepts of terminal controllers and to apply these developments to practical control systems. The Laplace transform method of solution of the aircraft differential equation of motion was introduced by this group, supplementing the time domain analyses previously employed. The frequency domain approach was found to be more meaningful to control system engineers, and it pointed the way to various simplifications in the control system development.

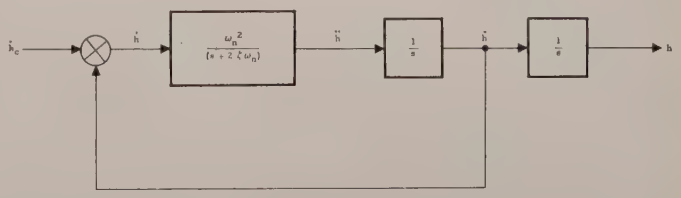
II. Prediction Equations for Landing Aircraft

The first step in the synthesis of terminal control systems is to develop prediction equations for the system to be controlled. These prediction equations represent the time solution of the system differential equations. The time solutions are written as a function of a fixed time in the future T by a time translation. The form of the prediction equations is a sum of products of the system state variables at the time t and weighting functions which are functions of the difference between the present terminal time t and the terminal time T. The exact definition of weighting functions used in this context will be deferred until later.

It is important to simplify the mathematical model of the system to be controlled. The degree of simplification must be consistent with the physical process involved. The complete differential equations representing a landing aircraft involve several simultaneous differential equations. The number of state variables is large and includes the following variables and their derivatives: pitch angle, altitude, velocity, flight path angle, and elevator deflection. The prediction equations required to define the system completely at the terminal time T would involve a large number of weighting functions and a large number of prediction equations. The method for obtaining terminal control equations for a generalized system with a number of simultaneous differential equations and a number of dependent variables is described in Appendix I. A straightforward application of the methods described in Appendix I to the complete differential equations describing the landing aircraft would result in a system too complex to be practical. Therefore, it is advisable to look for methods of simplifying the mathematical model of the landing aircraft before proceeding.

Often in the case of systems with many degrees of freedom the system can be characterized by some dominant mode (one or two characteristic roots) of the system. The system response to input commands and disturbances is largely dominanted by the characteristic roots of this mode. As a result the complex system can often be approximated by a simple second-order system.

It can be shown for the case of the landing aircraft that, if a high gain altitude rate loop is closed around the aircraft-autopilot combination, a simple mathematical model results. Fig. 1 shows a block diagram



Differential Equation

$$\frac{d^3h}{dt^3} + 2\zeta\omega_n\frac{d^2h}{dt^2} + \omega_n{}^2\frac{dh}{dt} = \omega_n{}^2\dot{h}_c(t)$$

where:
h = altitude
ζ = damping ratio
ω_n = angular natural frequency
$\dot{h}_c(t)$ = forcing function (altitude rate command)
t = independent variable (time)

Fig. 1—Altitude rate feedback loop.

which represents the dynamics of the landing aircraft with the altitude rate loop closed. This control loop ensures that the response from altitude rate command to altitude rate is that of a simple second-order system and can be considered uniform throughout the landing flare. The differential equation describing the block diagram is also shown in Fig. 1. This differential equation from altitude rate command to altitude is a third-order differential equation with constant coefficients as de-

[1] M. V. Matthews and C. W. Steeg, "Terminal Controller Synthesis," presented at Princeton Symposium on Non-Linear Control systems, March 26–27, 1956.

[2] R. C. Boonton, Jr., "Optimum design of final-value control systems," *Proc. Polytechnic Institute of Brooklyn Symp. on Non-Linear Control Systems*, 1956.

[3] R. V. Morris, "Design of a Motion Control System Subject to Constraints," M.S. thesis, Elec. Engrg. Dept., Mass. Inst. Tech., Cambridge, Mass.; 1956.

fined in Fig. 1. The prediction equations are obtained from this third order differential equation in Appendix II. The results of this development are summarized in (1) and (2):

$$h(T) = h(t) + h(t)[T - t] + \ddot{h}(t)F_{\ddot{h}}(T - t) + \ddot{h}_c(t)F_{\ddot{h}_c}(T - t) + h_\epsilon(t)F_{h_\epsilon}(T - t) \quad (1)$$

$$h(T) = h(t) + \ddot{h}(t)G_{\ddot{h}}(T - t) + \ddot{h}_c(t)G_{\ddot{h}_c}(T - t) + h_\epsilon(t)G_{h_\epsilon}(T - t). \quad (2)$$

Eq. (1) is the predicted value of the altitude at the terminal time T. This predicted value of altitude is in terms of the values of system variables existing at the present time t. These current variables are multiplied by weighting functions which are functions of the difference between the terminal time T and the present time t. These weighting functions depend upon the characteristics of the system shown in Fig. 1. Analytical expressions for these weighting functions are presented in Appendix II. Eq. (2) is a similar expression for the altitude rate at the terminal time T.

In Section III the prediction equations are mechanized to form a terminal controller for the landing system. As will be shown, the weighting functions in (1) and (2) are time-variable gains on various feedback signals.

In a practical landing system the terminal time T is determined during the system design. The automatic flare system begins to control the aircraft at some initial conditions of altitude and altitude rate. A nominal time is chosen for which the aircraft can be reasonably landed from the terminal conditions. There are certain physical constraints involved in determining the time required to land the aircraft in a satisfactory manner. This nominal time is best determined from analog computer simulation of the landing aircraft. This nominal time of flare then defines the terminal time T.

III. Controller Equations and Mechanization

The application of the prediction equations to control system synthesis is considered in this section. One of the quantities in the system differential equations must be the system control input. In the landing system equations of the previous section, the control input is the altitude acceleration command, $\ddot{h}_c(t)$. It is thus possible to determine from (1) the value of that control input which, if held constant during the extrapolation period (t to T), would result in the desired value of the controlled quantity at T. Since in the system described in the previous section, $\ddot{h}_c(t)$ is the control input, (1) may be solved for $\ddot{h}_c$. (It may be recalled that the form of (1) was derived using the simplifying assumption $\dddot{h}_c = 0$.) Thus the proper controlled parameter value at T ($h(T)_d$) can be obtained by establishing $\ddot{h}_c$ at this computed value and holding it there until time T. This mode of control shall be termed a type-A terminal controller. A block diagram depicting this control configuration is shown in Fig. 2. It will be noted that this method is not valid in the event $F_{\ddot{h}_c}$ passes through zero. [Note: Inasmuch as system variables have the argument (t), and weighting functions have the argument ($T-t$), these arguments may henceforth be omitted for brevity, *e.g.*, $\dot{h} \equiv \dot{h}(t)$ and $F_{\ddot{h}} \equiv F_{\ddot{h}}(T-t)$.]

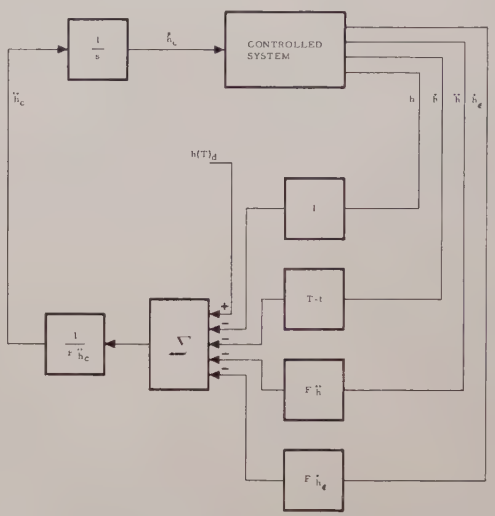


Fig. 2—Type-A terminal altitude controller.

Another approach to the control problem is to assume the control variable is zero and compare the desired value of the controlled parameter at T with that value determined from the prediction equation. This error can then be multiplied by a high gain to become the control variable. For a theoretical discussion of such a system carried out in detail see R. K. Smyth.[4] The authors show that the error signal decreases monotonically to zero in such a system and remains there unless the system is later disturbed. It is also demonstrated that if the system gain is increased to infinity, the system will remain stable, the error will be reduced to zero in minimum time, and there will be no error overshoot. In actual mechanizations, however, moderate gains perform adequately and simplify equipment requirements. A system utilizing such a control method will be referred to as a type-B controller. A block diagram showing this control technique is presented in Fig. 3.

Either of these two control techniques can be used to control one condition, *e.g.*, either h or $\dot{h}$ at $t-T$. Under these circumstances, such a system is referred to as a "one-condition terminal controller." However, it is further possible to control two terminal conditions simultaneously, in which case the system is termed a "two-condition terminal controller." The block diagram of Fig. 4 illustrates the mechanization of such a system, utilizing both type-A and type-B controllers, to control terminal values of both h and $\dot{h}$.

[4] R. K. Smyth, "Inertial Rate of Descent System," *Proc. IRE, ARS Spring Tech. Conf.*, Cincinnati, Ohio, pp. 1–14; April, 1960.

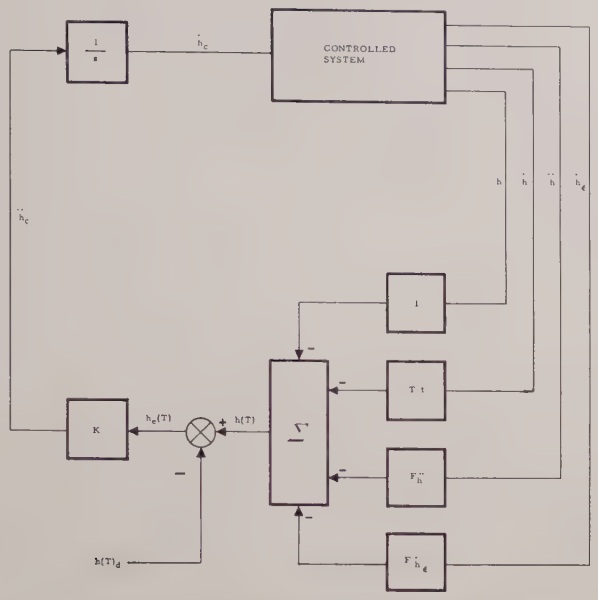


Fig. 3—Type-B terminal altitude controller.

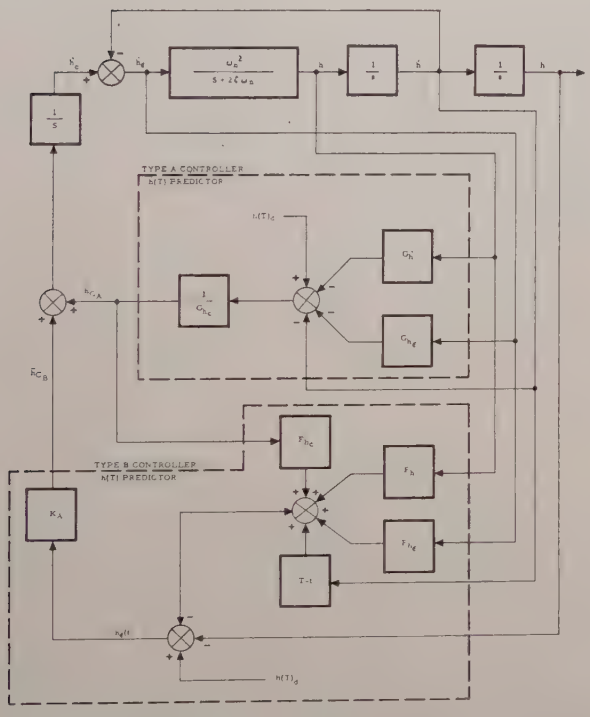


Fig. 4—Two-condition terminal controller landing system.

The system operation may be described as follows: The inner loop acts as a type-A controller, *i.e.*, it continually computes and feeds back the necessary $\ddot{h}_c$, which, if held constant until $t-T$, would result in obtaining the desired $\dot{h}(T)$.

$$\ddot{h}_{c_A} = \frac{1}{G_{\ddot{h}_c}} \left[\dot{h}(T)_d - h - \dot{h}G_{\dot{h}} - \dot{h}_\epsilon G_{\dot{h}_\epsilon}\right] \qquad (3)$$

where $\dot{h}(T)_d$ is the desired value of sinking rate at the terminal time. Closed around this loop is a type-B controller. The output of the $\dot{h}(T)$ predictor, which also uses $\dot{h}_c$ as an input, represents the error between the predicted and desired values of $h(T)$. This error $h_\epsilon(T)$ is fed back with a high gain (limited) to form $\ddot{h}_{c_B}$, thus tending to drive the error to zero.

$$\ddot{h}_{c_B} = K\left[h(T)_d - h - \dot{h}\left[T-t\right] - \ddot{h}F_{\ddot{h}} - \ddot{h}_{c_A}F_{\ddot{h}_c} + \dot{h}_\epsilon F_{\dot{h}_\epsilon}\right] \quad (4)$$

where $h(T)_d$ is the desired terminal altitude. Since this error goes rapidly to zero, the resulting configuration is that the control variable $\ddot{h}_c$ is made up entirely of the constant $\ddot{h}_{c_A}$ with $h_\epsilon(T)$ (and therefore $\ddot{h}_{cB}$) equal to zero. In the absence of additional disturbances, the system will then arrive at $t=T$ with the desired values of both h and $\dot{h}$. Any system disturbances will cause changes in $\ddot{h}_{cA}$ and $\ddot{h}_{cB}$, but the system will quickly come to rest at a constant value for $\ddot{h}_c(=\ddot{h}_{c_A})$, where both the $\dot{h}$ and h predictors are satisfied.

The value of feedback gain K is determined in practice from analog computer studies by gradually working down from high-gain values until further reduction results in discernible differences in system response. The resulting gains thus determined have been found actually to be only moderately high.

Up to this point, a formal synthesis procedure has been applied to the development of the terminal control landing system which has resulted in the configuration shown in Fig. 4. It is now possible to consider this configurarion as a feedback control system and combine the feedback paths in such a way that a significant simplification in the system results. The simplifications will be discussed in Section IV.

IV. Simplified Mechanizations

The purpose of this section of the paper is to consider two simplified versions of the two-condition terminal control system developed in Section III. The first simplification is in a form suitable for implementing a practical system. The second simplification is in a form which makes possible a stability analysis, thus giving additional insight into the type of control provided by the terminal controller.

The first simplification is developed in Appendix III and is illustrated in Fig. 5. This system has three feedback paths: altitude, altitude rate, and altitude rate error. The altitude rate and altitude rate error feedback signals are passed through weighting functions which can be considered as time variable gains. The altitude signal passes through a fixed gain. In addition, an altitude rate bias which represents the desired value of altitude rate at touchdown is passed through a weighting function and also biases the altitude rate feedback signal. The terminal control landing system in Fig. 5 can be readily implemented with two servo assemblies and three signal amplifiers. The three signal amplifiers are used for the necessary mixing and isolating functions indicated in Fig. 5. One of the two servo assemblies is used for the integration function indicated in Fig. 5. The second servo assembly is used for mechanizing the weighting functions.

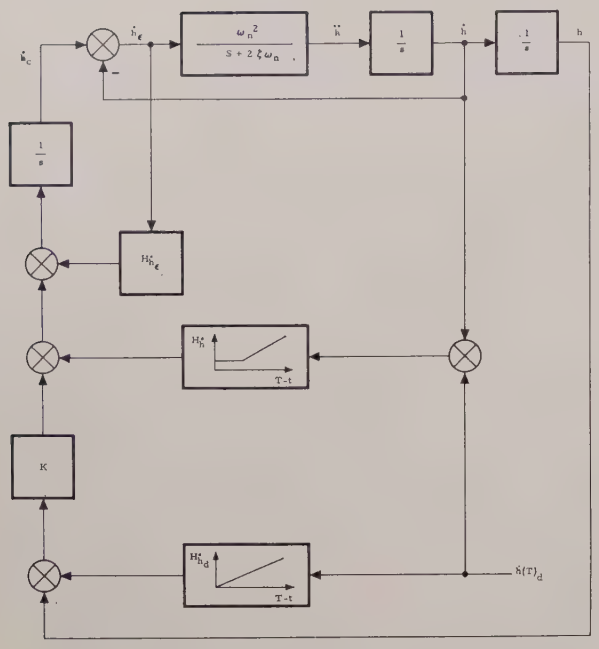


Fig. 5—Simplified terminal control landing system.

The weighting function servo assembly contains three function potentiometers. When the shaft is driven at a constant speed, the independent variable corresponding to the shaft position is $T-t$. This servo provides the "clock" function for the terminal controller.

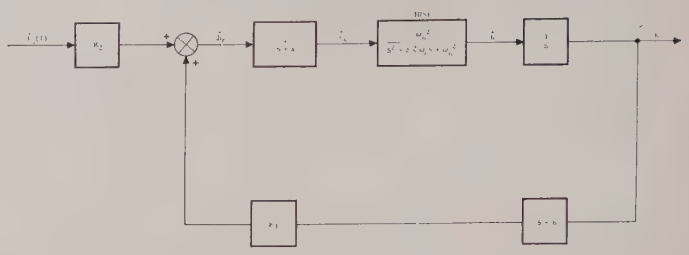
Only two weighting functions, $H_{\dot{h}_\epsilon}$ and $H_{\dot{h}}$, are nonlinear functions of $(T-t)$, requiring special shaping. The function $\dot{h}(T)_d\ [-K(T-t)]$ is formed by a linear potentiometer on the shaft, having a constant input. The input $\dot{h}_\epsilon$ to the $H_{\dot{h}_\epsilon}$ potentiometer is taken directly from the output of a summing amplifier. Biased altitude rate $(\dot{h}-\dot{h}(T)_d)$ is derived from an altitude rate sensing instrument (preferably a radar altimeter-monitored inertial sensor[5]), and the altitude h is taken from a radar altimeter. These three potentiometer outputs are summed, together with $[-K]h$ into the input of the $\dot{h}_c$ integrator, which may be an electronic integrator or an electro-mechanical servo with rate feedback. The difference between this altitude rate command and $\dot{h}$ (derived as above) is then available at a summing amplifier output to be used as an input command to the autopilot inner pitch rate loop.

It is thus possible to provide a simple landing system mechanization and yet maintain the inherent advantages of the terminal control concept.

A second simplification developed in Appendix IV yields a block diagram of the terminal controller as illustrated in Fig. 6. This configuration results when the various feedback signals in Fig. 4 are combined using the numerical values existing at some particular fixed time. With this simplification, it will be noted that the terminal controller can be viewed as a conventional control system with various "transfer functions" in the feedforward and feedback loop. The quotations on transfer function are used to remind the reader that the transfer functions obtained are valid only at a specific time and only in a restricted sense. The gains and time constants of the terminal control "transfer functions" are functions of time. A numerical tabulation of these parameters presented in Table I of Appendix IV shows that these parameters vary slowly with time. This slow variation leads one to consider a stability analysis in which time is "frozen," or the parameters a, b, K_1 and K_2 are considered constant.

Such a stability analysis is illustrated by the root locus in Fig. 7. The root locus indicates that the damping of the altitude roots is reduced somewhat by the action of the terminal controller. However, the dom-

[5] E. R. Buxton and E. W. Velander, "Landing System for Commercial Aircraft," presented to the IATA, Montreal, Quebec, Can.; April 6, 1959.



Note: K_1, K_2, a, and b are functions of $T-t$

Fig. 6—Terminal controller as feedback system with time-varying gains and time constants.

TABLE I

h_c TRANSFER FUNCTIONS AT SELECTED INTERVALS*

$T-t = 20$ sec

$$\dot{h}_c = \frac{-.8(s+2.5)}{(s+20.7)}\,h(s) - \frac{18.5}{s+20.7}\,\dot{h}_D(T)$$

$T-t = 16$ sec

$$\dot{h}_c = \frac{1.05(s+1.92)}{(s+17)}\,h(s) - \frac{14}{s+17}\,\dot{h}_D(T)$$

$T-t = 12$ sec

$$\dot{h}_c = \frac{1.28(s+1.55)}{s+13.3}\,h(s) - \frac{9.5}{s+13.3}\,\dot{h}_D(T)$$

$T-t = 8$ sec

$$\dot{h}_c = \frac{1.54(s+1.3)}{(s+9.5)}\,h(s) - \frac{5}{s+9.5}\,\dot{h}_D(T)$$

$T-t = 4$ sec

$$\dot{h}_c = \frac{1.838(s+1.09)}{(s+5.881)}\,h(s) - \frac{.281}{s+5.881}\,\dot{h}_D(T)$$

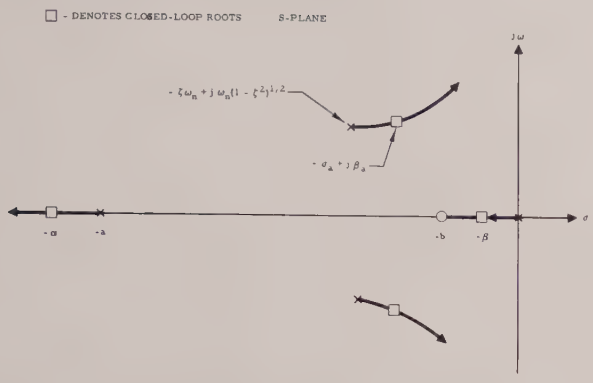
$T-t = 2$ sec

$$\dot{h}_c = \frac{2.1(s+.953)}{(s+2.7)}\,h(s) - \frac{.8}{s+2.7}\,\dot{h}_D(T)$$

$T-t = 1$ sec

$$\dot{h}_c = \frac{2.6(s+.77)}{(s+.4)}\,h(s) - \frac{1}{s+.4}\,\dot{h}_D(T)$$

* The coefficients of the following $\dot{h}_c$ transfer functions were taken from the numerical values of the weighting functions of Fig. 4. Curves 1 through 4 were plotted for these gain values as functions of time. These expressions are the numerical values for (45).



Characteristic equation:

$$\left[\frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}\right]\frac{(-K_1)(s+b)}{s(s+a)} = -1$$

Closed loop:

$$\frac{h}{h_d(\tau)}(s) = \frac{\omega_n^2}{(s+\alpha)(s+\beta)(s+\sigma_a \pm j\beta_a)}$$

Fig. 7—Stability of terminal controller at fixed time.

inant response of the system will result from the real root located at $-\beta$, giving rise to a decaying exponential type response. The time constant of this exponential is growing smaller as the terminal time is approached.

It should be remembered that the point stability analysis suggested in this example is quite analogous to the point stability analyses commonly conducted on a missile booster for which the differential equations describing the booster motion have time-varying coefficients. In a like manner as in this analysis of the terminal controller, the coefficients of the booster differential equations are assumed fixed at the point at which the stability is considered.

V. Test Results

The basic concept of the terminal control landing systems discussed in this paper has been verified by extensive analog computer simulations and two flight test programs. The first flight test program involved a TF-100F and was conducted during the winter of 1959 and 1960. The TF-100F configuration was similar to that shown in Fig. 4. This program resulted in a number of successful automatic landing flares and tested the concept of manual instrument flares utilizing the pilot to null the pitch bar of the attitude displays indicator during the flare maneuver.

A simplified version of the autoflare system similar to that shown in Fig. 5 was successfully flight-tested in a TF-102A during the summer and early fall of 1960. A number of successful automatic landings were accomplished in this program at several different airports in the Los Angeles area.

These landings were accomplished with the pilot unable to see outside the cockpit. The terminal control landings were considered to be satisfactory from a pilot's and passenger's point of view. The touchdown sinking rate was 2 feet/sec or less in most cases. The aircraft touched down within 50 feet longitudinally of the same point on the runway under the test conditions.

Two very important characteristics of the terminal control landing system have been observed in the analog and flight test results: 1) the terminal control concept used makes the landing performance singularly insensitive to off-nominal initial conditions, gain variations and gust disturbances, and 2) the system is considerably easier to land in the manual instrument mode than the previous landing systems which have been investigated, such as the exponential flare type of landing system and the programmed path type landing flare computer. The superiority of the terminal control system is particularly pronounced in the presence of gust disturbances. This characteristic allows a qualified instrument pilot to make consistently good landings even under adverse conditions.

Fig. 8 illustrates the effect of off-nominal initial conditions on the landing flare trajectory. The nominal sinking rate for the TF-102A at flare trigger is 15feet/sec (900 feet per minute). Fig. 8(a) shows the altitude and altitude rate traces as a function of time for an analog computer run. Fig. 8(b) and (c) show the effects of plus and minus 50 per cent variation from the nominal sinking rate. It will be noted that very little change in the flare path shape occurs. The desired terminal conditions of 2 feet/sec altitude are met for this large variation in initial sinking rate. The time at which touchdown occurs varies by approximately $1\frac{1}{2}$ seconds from the nominal value of 13 seconds. The variation represents a range variation of approximately 350 feet. The initial altitude rate variations are considerably greater than is to be expected under normal circumstances.

Fig. 9 shows a comparison of a typical analog computer simulation with flight test results of the terminal control landing system. Even though the initial sinking rates of the two traces were not exactly the same, the shape of the flares compared very closely. In general a close agreement was found between the flight test results and the analog computer simulation of the autoflare system.

According to the theory developed in Appendix II, non-nominal aircraft velocity histories affect the flare only inasmuch as the aircraft altitude rate response is changed. As this response deviation is generally slight, no appreciable flare differences are anticipated resulting from these velocity dispersions.

By way of confirmation of this conclusion, analog computer simulations have indicated that longitudinal velocity variations prior to landing flare initiation have no significant effect on the landing flare trajectory.

VI. Conclusions and Summary

In this paper, a procedure is described which can be used to synthesize a terminal controller when the differential equations of a system are known. This technique is illustrated for the case of a landing aircraft with an altitude rate autopilot.

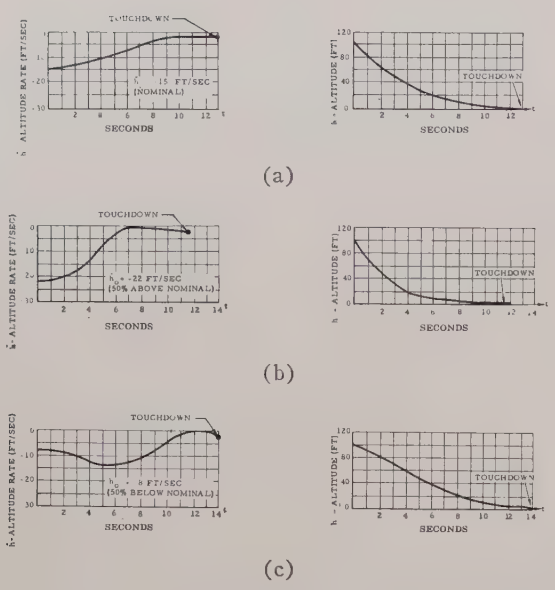


Fig. 8—Altititude and altitude rate trajectories for several values of initial altitude rate.

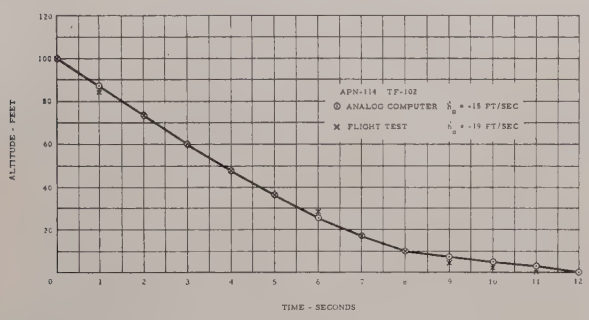


Fig. 9—Comparison of analog computer simulation and flight test results of terminal control landing system.

The procedure gives rise to a controller which can be simply implemented using standard servo hardware. The net result of the synthesis procedure is a feedback controller with gains which are varied as a function of time in a desirable manner.

It is suggested that the terminal control technique could be applied to many present-day and future control problems in which the system differential equations can be written. There are a number of space applications such as a soft lunar landing and orbital rendezvous problems for which terminal controllers may find useful applications.

Appendix I

Terminal Controller Equations for a Generalized System

As pointed out in the introduction, the value of any system state variables at any specified time in the future can be determined from a knowledge of the differential equations which describe the motion of the system and the present values of all the system state variables.

This simple statement is the essence of the prediction equations, which are nothing more than time solutions of differential equations, based upon initial conditions taken always at the present. The time solution of these differential equations may be obtained in any manner desired. In the following discussion, solutions for a generalized system are obtained by the Laplace transform method, which is well suited for the solution of simultaneous sets of system differential equations with known initial conditions. Consider a system described by the two differential equations

$$a_n x^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_o x + b_m y^{(m)} + b_{m-1}y^{(m-1)} + \cdots + b_o y = F(t) \quad (5)$$

$$c_n x^{(n)} + c_{n-1}x^{(n-1)} + \cdots + c_o x + g_m y^{(m)} + g_{m-1}y^{(m-1)} + \cdots + g_o y = KF(t) \quad (6)$$

where $x^{(n)}$ denotes the nth derivative with respect to time $d^n x/dt^n$. Taking the Laplace transform of these equations yields

$$\begin{aligned}&(a_n s^n + a_{n-1}s^{n-1} + \cdots + a_o)X(s) \\ &+ (b_m s^m + b_{m-1}s^{m-1} + \cdots + b_o)Y(s) = F(s) \\ &+ (a_n s^{n-1} + a_{n-1}s^{n-2} + \cdots + a_1)x(o) \\ &+ (a_n s^{n-2} + a_{n-1}s^{n-3} + \cdots + a_2)x^{(1)}(o) + \cdots \\ &+ a_n x^{(n-1)}(o) \\ &+ (b_m s^{m-1} + b_{m-1}s^{m-2} + \cdots + b_1)y(o) \\ &+ (b_m s^{n-2} + b_{m-1}s^{m-3} + \cdots + b_2)y^{(1)}(o) + \cdots \\ &+ b_m y^{(m-1)}(o) \end{aligned} \quad (7)$$

and

$$\begin{aligned}&(c_n s^n + c_{n-1}s^{n-1} + \cdots + c_o)X(s) \\ &+ (g_m s_m + g_{m-1}s^{m-1} + \cdots + g_o)Y(s) = KF(s) \\ &+ (c_n s^{n-1} + c_{n-1}x^{n-2} + \cdots + c_1)x(o) \\ &+ (c_n s^{n-2} + c_{n-1}s^{n-3} + \cdots + c_2)x^{(1)}(o) + \cdots \\ &+ c_m x^{(n-1)})o) + (g_m s^{m-1} + g_{m-1}s^{m-2} + \cdots + g_1)y(o) \\ &+ (g_m s^{m-2} + g_{m-1}s^{m-3} + \cdots + g_2)y^{(1)}(o) + \cdots \\ &+ g_m y^{(m-1)}(o) \end{aligned} \quad (8)$$

where $x^n(o)$ denotes $d^n x/dt^n\big|_{t=0}$, *i.e.*, the nth time derivative of x evaluated at $t=0$. These equations (7) and (8) can then be solved as algebraic equations for $X(s)$ and $Y(s)$, from which desired time solutions can be written by taking the inverse Laplace transform. For instance,

$$\begin{aligned}x(t) = {} & x^{(n-1)}(o)H_{n-1}(t) + x^{(n-2)}(o)H_{n-2}(t) + \cdots \\ & + x(o)H_o(t) + y^{(m-1)}(o)J_{m-1}(t) \\ & + y^{(m-2)}(o)J_{m-2}(t) + \cdots + y(o)J_1(t) \\ & + \int_0^t F(\lambda)I(T - t - \lambda)d\lambda \end{aligned} \quad (9)$$

where H, J, and I are the so-called weighting functions which are (in this case) found as follows:

$$H_{n-1}(t) = \mathfrak{L}^{-1}(H_{n-1}(s)) = \mathfrak{L}^{-1}\left(\frac{\Delta^{X}_{x^{(n-1)}}(s)}{\Delta(s)}\right) \tag{10}$$

where $\Delta(s)$ is the system characteristic determinant and $\Delta^{X}_{x^{(n-1)}}(s)$ is the system determinant with the $x(s)$ column replaced by the $x^{(n-1)}(o)$ coefficients.

Similarly

$$J_{m-1}(t) = L^{-1}\left(\frac{\Delta^{X}_{y^{(m-1)}}(s)}{\Delta(s)}\right) \tag{11}$$

and

$$I(t) = L^{-1}\left(\frac{\Delta^{X}_{F}(s)}{\Delta(s)}\right). \tag{18}$$

Of course, a similar expression can be obtained for $y(t)$.

Eq. (10) is the desired expression for x as a function of time and the initial conditions of the various system parameters. At this point it is convenient to generalize the extrapolation relationship (9) to express the predicted value of a desired parameter at a future time T by observation of the current values (initial conditions) of the system parameters at any prior time t. This can be accomplished merely by a shift of origin of (9) and a redesignation of the time variable in such a way that the resulting expression is meaningful in terms of a fixed time reference. If the time at which the value of x is desired is T and the time at which the various parameter values (initial conditions) are known is t, then the period of extrapolation becomes $T=t$ and the prediction equation becomes

$$\begin{aligned} x(T) = {} & x^{(n-1)}(t)H_{n-1}(T-t) \\ & + x^{(n-2)}(t)H_{n-2}(T-t) + \cdots + x(t)H_o(T-t) \\ & + y^{(m-1)}(t)J_{m-1}(T-t) \\ & + y^{(m-2)}(t)J_{m-2}(T-t) + \cdots + y(t)J_o(T-t) \\ & + \int_t^T F(\lambda)I(T-t-\lambda)d\lambda. \end{aligned} \tag{19}$$

It is then seen that in a linear system the value of any parameter x existing at future time T may be predicted based upon values of the system quantities x and y and their derivatives (up to an order one less than the highest-order derivative of the particular quantity appearing in the differential equations) at any time t prior to T. In addition, a knowledge of the future behavior of the forcing function F must be had. The convolution integral is avoided if the transform of F can be taken [(7) and (8)]. In terminal control systems F is usually taken to be zero, a constant, or a ramp which simplifies the prediction equation.

It is seen that the weighting functions such as $H_{n-1}(T-t)$ and $J_{m-1}(T-t)$ are the time solution contributions to x produced by unit values at time t of the associated parameters, *e.g.*, $H_{n-1}(T-t)$ is the amount by which x changes between t and T due to a unit value of $x^{(n-1)}$ at time t. Because of linearity, the total prediction $x(T)$ is made up of the sum of all such prediction components. As outlined above, these weighting functions can be calculated mathematically, but they can also be determined directly on an analog computer. The definition of these weighting functions suggests that, once the system equations are set up on the computer, a given weighting function, say $H_{n-1}(T-t)$ may be determined simply by starting the problem with all the parameters (the various quantities and their $p-1$ derivatives, where p is the highest-order derivative of that variable appearing in the original equations) set to zero except $x^{(n-1)}$, which is set to unity.[6] The weighting function is then the resulting x output. For similar physical reasons, it is apparent that the problem solution can depend only upon $F(t)$ and the initial values of all the integrators in the problem. Thus the analog computer can be used to determine what weighting functions are required and their actual shapes. Since most complex problems will be analyzed on the computer to evaluate system performance, if nothing more, it is most desirable from the standpoints of time and accuracy to obtain also the required weighting functions in this manner.

Appendix II

Development of Landing Aircraft Prediction Equations

In this appendix the altitude and altitude rate prediction equations for the simplified model of the landing aircraft (Fig. 1) are developed.

The third-order differential equation in Fig. 1 may be solved by taking the Laplace transform, including the initial conditions in the manner of Appendix I:

$$\begin{aligned} & [s^3H(s) - s^2h(o) - sh(o) - \ddot{h}(o)] \\ & \quad + 2\zeta\omega_n[s^2H(s) - sh(o) - h(o)] \\ & \quad + \omega_n^2[sH(s) - h(o)] = \omega_n^2H_c(s), \end{aligned} \tag{20}$$

where

s = complex frequency variable
$H(s)$ = Laplace transform of $h(t)$
$h(o)$ = initial condition of altitude
$\dot{h}(o)$ = initial condition of altitude rate
$\ddot{h}(o)$ = initial condition of altitude acceleration
$H_c(s)$ = Laplace transform of $h_c(t)$.

[6] Since these parameters are invariably represented by integrator outputs, this requires merely that all the integrators (except one) be set to begin the problem with no initial conditions applied.

This equation may be solved for $H(s)$ with the following result:

$$H(s) = h(o)\frac{1}{s} + h(o)\left[\frac{s + 2\zeta\omega_n}{s(s^2 + 2\zeta\omega_n + \omega_n^2)}\right] + \ddot{h}(o)\left[\frac{1}{s(s^2 + 2\zeta\omega_n + \omega_n^2)}\right] + \dot{H}_c(s)\left[\frac{\omega_n^2}{s(s^2 + 2\zeta\omega_n s + \omega_n^2}\right]. \quad (21)$$

In taking the inverse Laplace transform of (21) to obtain $h(t)$, no difficulty is encountered except with the last term. In order to complete the inverse transformation, it is necessary to know the form of $\dot{H}_c(s)$. Justification for assumption of a simple form, namely a step for $\dot{h}_c(t)$, is provided in Appendix I. A more desirable form of $\dot{h}_c$ for a landing system is a ramp input, which corresponds to a step input into an integrator preceding the $\dot{h}_c(t)$ point of Fig. 1. A ramp input ensures a smoother control than the step input. For the ramp input

$$h_c(t) = h_c(o) + \ddot{h}_c(o)t. \quad (22)$$

The Laplace transform is

$$H_c(s) = h_c(o)\left[\frac{1}{s}\right] + \ddot{h}_c(o)\left[\frac{1}{s^2}\right]. \quad (23)$$

Substituting (23) into (21) yields

$$H(s) = h(o)\frac{1}{s} + \dot{h}(o)\left[\frac{s + 2\zeta\omega_n}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right] + \ddot{h}(o)\left[\frac{1}{s(s^2 + 2\zeta\omega_n s + \omega_n)^2}\right] + h_c(o)\left[\frac{\omega_n^2}{s^2(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right] + \ddot{h}_c(o)\left[\frac{\omega_n^2}{s^3(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right]. \quad (24)$$

Taking the inverse transform of (24), the following time equation is obtained:

$$h(t) = h(o)F_h(t) + \dot{h}(o)F_{\dot{h}}(t) + \ddot{h}(o)F_{\ddot{h}}(t) + h_c(o)F_{h_c}(t) + \ddot{h}_c(o)F_{\ddot{h}_c}(t) \quad (25)$$

where the F's are the altitude weighting functions defined in Fig. 10. Fig. 10 also shows sketches of the shape of the weighting functions vs time.

Eq. (25) represents the value of altitude h that exists at the present time (t) as a consequence of certain conditions $h(o)$, $\dot{h}(o)$, $\dot{h}_c(o)$, $\ddot{h}_c(o)$ having existed at some previous time t seconds ago. An equation for $\dot{h}(t)$ can be derived in a similar manner, or it can be obtained by directly differentiating (25), keeping in mind that only the F's are time-varying. In either case an equation for $\dot{h}(t)$ results as follows:

$$\dot{h}(t) = \dot{h}(o)G_{\dot{h}}(t) + \ddot{h}(o)G_{\ddot{h}}(t) + \ddot{h}_c G_{\ddot{h}_c}(t) + h_c(o)G_{\dot{h}_c}(t) \quad (26)$$

where the G's are the time derivatives of the corresponding F's and represent another set of weighting functions which are defined together with the weighting function shapes in Fig. 11.

A substitution made at this point will result in a simplification of two of the shapes which must be mechanized in order to form the prediction equations.

Notice that $\dot{h}_c$, $\dot{h}_\epsilon$ and $\dot{h}$ are related as follows (Fig. 1):

$$h_c = h + h_\epsilon. \quad (27)$$

If this equation is substituted into (25) and (26) to eliminate $\dot{h}_c$, and terms are recollected, the following equations result:

$$h(t) = h(o)F_h(t) + \dot{h}(o)[F_{\dot{h}}(t) + F_{\dot{h}_c}(t)] + \ddot{h}(o)F_{\ddot{h}}(t) + \ddot{h}_c(o)F_{\ddot{h}_c}(t) + h_\epsilon(o)F_{\dot{h}_c}(t) \quad (28)$$

and

$$h(t) = h(o)[G_{\dot{h}}(t) + G_{\dot{h}_c}(t)] + \ddot{h}(o)G_{\ddot{h}}(t) + \ddot{h}_c(o)G_{\ddot{h}_c}(t) + h_\epsilon(o)G_{\dot{h}_c}(t). \quad (29)$$

Now the weighting functions that are collected together may be examined.

$$F_{\dot{h}}(t) + F_{\dot{h}_c}(t) = \mathcal{L}^{-1}\left\{\frac{s + 2\zeta\omega_n}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)} + \frac{\omega_n^2}{s^2(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\} = \mathcal{L}^{-1}\left\{\frac{s^2 + 2\zeta\omega_n s + \omega_n^2}{s^2(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\} = \mathcal{L}^{-1}\left\{\frac{1}{s^2}\right\} = t \quad \text{(a ramp)}. \quad (30)$$

Therefore,

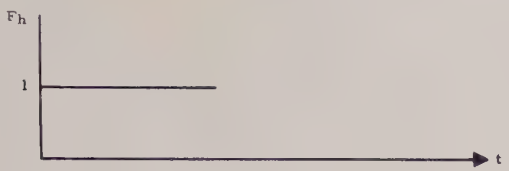
$$F_{\dot{h}}(t) + F_{\dot{h}_c}(t) = [t]. \quad (31)$$

(The brackets are used to avoid confusion with the functional notation.) In similar manner,

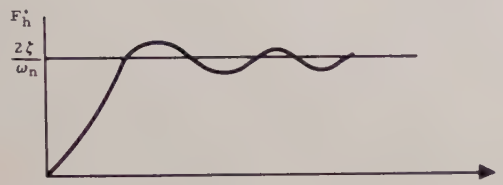
$$G_{\dot{h}}(t) + G_{\dot{h}_c}(t) = \mathcal{L}^{-1}\left\{\frac{s + 2\zeta\omega_n}{(s^2 + 2\zeta\omega_n s + \omega_n^2)} + \frac{\omega_n^2}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\} = \mathcal{L}^{-1}\left\{\frac{s^2 + 2\zeta\omega_n s + \omega_n^2}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\} = \mathcal{L}^{-1}\left\{\frac{1}{s}\right\} = 1 \quad \text{(unity)}. \quad (32)$$

Therefore:

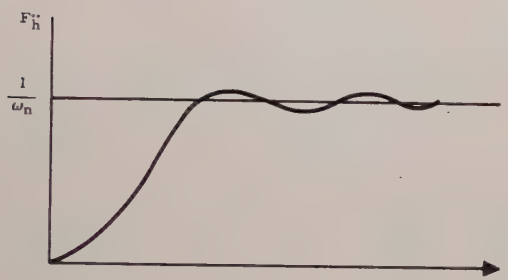
$$G_{\dot{h}}(t) + G_{\dot{h}_c}(t) = 1. \quad (33)$$



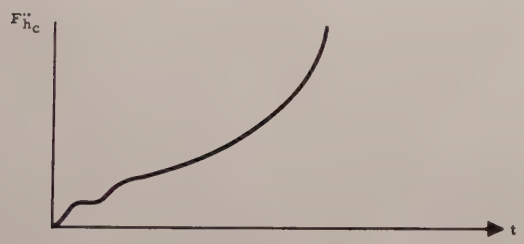
$$F_h(t) = \mathcal{L}^{-1}\left\{\frac{1}{s}\right\} = 1$$



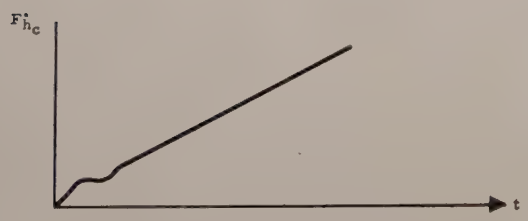
$$F_h^{\cdot}(t) = \mathcal{L}^{-1}\left\{\frac{s + 2\zeta\omega_n}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\}$$



$$F_h^{\cdot\cdot}(t) = \mathcal{L}^{-1}\left\{\frac{1}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\}$$

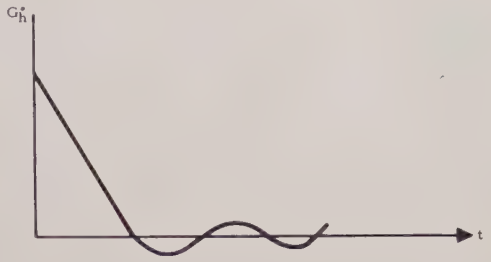


$$F_{h_c}^{\cdot\cdot}(t) = \mathcal{L}^{-1}\left\{\frac{\omega_n^2}{s^3(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\}$$

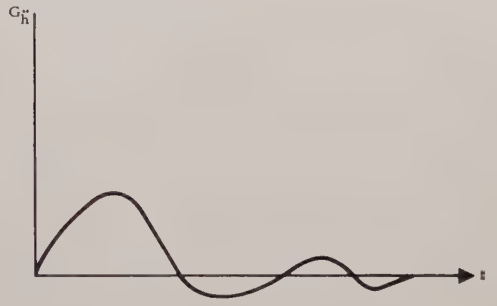


$$F_{h_c}^{\cdot}(t) = \mathcal{L}^{-1}\left\{\frac{\omega_n^2}{s^2(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\}$$

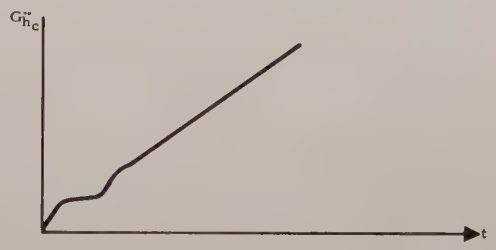
Fig. 10—Altitude weighting functions.



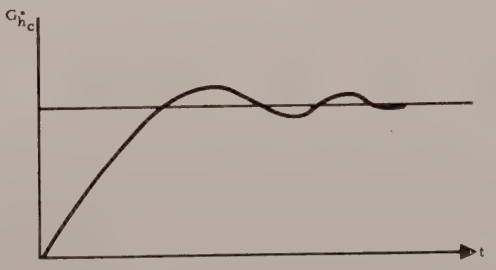
$$G_h^{\cdot}(t) = \mathcal{L}^{-1}\left\{\frac{s + 2\zeta\omega_n}{s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\}$$



$$G_h^{\cdot\cdot}(t) = \mathcal{L}^{-1}\left\{\frac{1}{s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\}$$



$$G_{h_c}^{\cdot\cdot}(t) = \mathcal{L}^{-1}\left\{\frac{\omega_n^2}{s^2(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\}$$



$$G_{h_c}^{\cdot}(t) = \mathcal{L}^{-1}\left\{\frac{\omega_n^2}{s(s^2 + 2\zeta\omega_n s + \omega_n^2)}\right\}$$

Fig. 11—Altitude rate weighting functions.

Substituting (32) and (34) into (29) and (30) and hereafter defining $F_{\dot{h}_\epsilon}=F_{\dot{h}_c}$ and $G_{\dot{h}_\epsilon}=G_{\dot{h}_c}$, the following equations result:

$$h(t) = h(o) + \dot{h}(o)[t] + \ddot{h}(o)F_{\ddot{h}}(t) + \ddot{h}_c(o)F_{\ddot{h}_c}(t) + \dot{h}_\epsilon(o)F_{\dot{h}_\epsilon}(t) \quad (34)$$

$$\dot{h}(t) = \dot{h}(o) + \ddot{h}(o)G_{\ddot{h}}(t) + \ddot{h}_c(o)G_{\ddot{h}_c}(t) + \dot{h}_\epsilon(o)G_{\dot{h}_\epsilon}(t). \quad (35)$$

Eqs. (34) and (35) can be generalized by considering the present conditions the initial conditions [argument (t) rather than (o)], thus requiring an argument $(T-t)$ for the variable terms when considering a general time T. This shift of time reference yields

$$h(T) = h(t) + \dot{h}(t)[T-t] + \ddot{h}(t)F_{\ddot{h}}(T-t) + \ddot{h}_c(t)F_{\ddot{h}_c}(T-t) + \dot{h}_\epsilon(t)F_{\dot{h}_\epsilon}(T-t) \quad (36)$$

$$\dot{h}(T) = \dot{h}(t) + \ddot{h}(t)G_{\ddot{h}}(T-t) + \ddot{h}_c(t)G_{\ddot{h}_c}(T-t) + \dot{h}_\epsilon(t)G_{\dot{h}_\epsilon}(T-t) \quad (37)$$

Eqs. (36) and (37), which are the same as (20) and (21), are the so-called prediction equations, which extrapolate, by means of differential equation solution, the existing conditions to the conditions at future time T. It is to be recalled that this extrapolation is predicated on the assumption of the constancy of the input rate $\ddot{h}_c$ from present time t until terminal time T.

Appendix III

Simplification of Terminal Controller Mechanization

In this appendix the simplification of the terminal controller mechanization suitable for implementation (described in Section IV) is developed.

This simplification begins with the two-condition terminal controller illustrated in Fig. 4 and combines feedback paths by algebraic manipulation of the weighting functions.

The expression for altitude acceleration command $\ddot{h}_c$ can be written as

$$\ddot{h}_c = \dot{h}_\epsilon\left[-\frac{G_{\dot{h}}}{G_{\ddot{h}_c}}(1-KF_{\ddot{h}_c})-KF_{\dot{h}_\epsilon}\right] + \cdots + \ddot{h}\left[-\frac{G_{\ddot{h}}}{G_{\dot{h}_c}}(1-KF_{\ddot{h}_c})-KF_{\ddot{h}}\right] + \dot{h}\left[-\frac{1}{G_{\ddot{h}_c}}(1-KF_{\ddot{h}_c})-K[T-t]\right] + \dot{h}(T)_d\left[+\frac{1}{G_{\ddot{h}_c}}(1-KF_{\ddot{h}_c})\right] - h[K] + h(T)_d[K]. \quad (38)$$

This technique of combination is valid for finite values of gain K and sufficiently high limits on $\ddot{h}_{c_a}$ and $\ddot{h}_{c_b}$. It has been determined from simulation studies that the effect of $\ddot{h}$ is usually small since it has relatively small values for the weighting functions and can be neglected. The last term of Eq. (38), the desired terminal altitude, is zero. Considering these points, (38) may be arranged as

$$\ddot{h}_c = \dot{h}_\epsilon\left[-\frac{G_{\dot{h}_\epsilon}}{G_{\ddot{h}_c}}(1-KF_{\ddot{h}_c})-KF_{\dot{h}_\epsilon}\right] + (h-\dot{h}(T)_d)\left[-\frac{1}{G_{\ddot{h}_c}}(1-KF_{\ddot{h}_c})-K[T-t]\right] + \dot{h}(T)_d[-K[T-t]] - h[K] \quad (39)$$

or

$$\ddot{h}_c = \dot{h}_\epsilon[H_{\dot{h}_\epsilon}] + (\dot{h}-\dot{h}(T)_d)[H_{\dot{h}}] + \dot{h}(T)_d[-K[T-t]] + h[-K] \quad (40)$$

thereby defining new composite weighting functions $H_{\dot{h}_\epsilon}(T-t)$ and $H_{\dot{h}}(T-t)$. The landing system mechanization, thus simplified, is that shown in Fig. 5.

It may be noted that all F's and G's appearing in (38) approach zero as $(T-t)\to 0$. Accordingly the $(\dot{h}-\dot{h}(T)_d)$ term becomes important and quite large. It is therefore reasonable to cause other weighting functions to go to zero and to limit $H_{\dot{h}}$. This effectively establishes a control system designed to hold $\dot{h}=\dot{h}(T)_d$ near the ground as time runs out, which corresponds to the control philosophy expected at this point of the flare.

Appendix IV

Terminal Controller as Servo with Time-Varying Parameters

This appendix provides a simplification which is obtained by combining the feedback paths of the terminal controller in Fig. 4. This approach obtains an equivalent transfer function for the terminal controller assuming that time is fixed or "frozen." The following equations can be written from inspection of Fig. 4:

$$s\dot{h}_c = \dot{h}_\epsilon\left(-\frac{G_{\dot{h}_\epsilon}}{G_{\ddot{h}_c}} - K_AF_{\dot{h}_\epsilon} + K_A\frac{G_{\dot{h}_\epsilon}F_{\ddot{h}_c}}{G_{\ddot{h}_c}}\right) + \dot{h}\left(-\frac{1}{G_{\dot{h}_c}} - K_A(T-t_i) + K_A\frac{F_{\ddot{h}_c}}{G_{\ddot{h}_c}}\right) - hK_A + \dot{h}_D(T)\left(-K_A\frac{F_{\ddot{h}_c}}{G_{\ddot{h}_c}} + \frac{1}{G_{\ddot{h}_c}}\right) \quad (41)$$

but

$$h_\epsilon = h_c - h. \quad (42)$$

Combining (41) and (42) to eliminate $\dot{h}_\epsilon$, the following is obtained:

$$s\dot{h}_c = \dot{h}_c\left(-\frac{G_{\dot{h}_\epsilon}}{G_{\ddot{h}_c}} - K_A F_{\dot{h}_\epsilon} + K_A\frac{G_{\dot{h}_c}F_{\ddot{h}_c}}{G_{\ddot{h}_c}}\right) + \dot{h}\left(-\frac{1}{G_{\ddot{h}_c}} - K_A(T-t_i)\right) + K_A\frac{F_{\ddot{h}_c}}{G_{\ddot{h}_c}} + \frac{G_{\dot{h}_\epsilon}}{G_{\ddot{h}_c}} + K_A F_{\dot{h}_\epsilon} - K_A\frac{G_{\dot{h}_\epsilon}F_{\ddot{h}_c}}{G_{\ddot{h}_c}}\right) - hK_A + \dot{h}_D(T)\frac{1}{G_{\ddot{h}_c}}(1 - K_A F_{\ddot{h}_c}). \quad (43)$$

Collecting terms in $\dot{h}_c(s)$ on the left, obtain

$$\dot{h}_c(s)\left[s + \frac{G_{\dot{h}}}{G_{\ddot{h}_c}}(1 - K_A F_{\ddot{h}_c}) + K_A F_{\dot{h}_\epsilon}\right] = -h(s)\left[K_A + s\left(\frac{K_A F_{\ddot{h}_c} + G_{\dot{h}_\epsilon}(1 - K_A F_{\ddot{h}_c}) - 1}{G_{\ddot{h}_c}} + K_A F_{\dot{h}} - K_A(T - t_i)\right)\right] + \dot{h}_D(T)\left[\frac{1}{G_{\ddot{h}_c}}(1 - K_A F_{\ddot{h}_c})\right]. \quad (44)$$

Eq. (44) may be manipulated easily into the form

$$h_c(s) = +h(s)\frac{K_1(s+b)}{(s+a)} + h_D(T)\frac{K_2}{(s+a)} \quad (45)$$

where

$$a = \frac{G_{\dot{h}}}{G_{\ddot{h}_c}}(1 - K_A F_{\ddot{h}_c}) + K_A F_{\dot{h}}.$$

$$b = \frac{1}{\dfrac{F_{\ddot{h}_c} + \dfrac{G_{\dot{h}}}{K_A}(1 - K_A F_{\ddot{h}_c}) - \dfrac{1}{K_A}}{G_{\ddot{h}_c}} + F_{\dot{h}} - (T - t_i)}$$

$$K_1 = \frac{+K_A F_{\ddot{h}_c} + G_{\dot{h}}(1 - K_A F_{\ddot{h}_c}) - 1}{G_{\ddot{h}c}} + [F_{\dot{h}} - (T - t_i)]K_A$$

$$K_2 = +\frac{1}{G_{\ddot{h}_c}}(1 - K_A F_{\ddot{h}_c}).$$

Using (45) as the representation for the terminal controller, a simplified block diagram of the landing system may be drawn as illustrated in Fig. 6.

In Table I, a tabulation of the expression for $\dot{h}_c(s)$ at seven different values of $T-t$ from 20 to 1 second is presented for the terminal controller representation of Fig. 6. This tabulation should give the reader a feeling for the magnitude of a, b, K_1 and K_2, and also how these parameters vary as a function of time. It will be noted from the tabulation that the parameters vary relatively slowly with time.

Acknowledgment

The authors wish to acknowledge the contributions to terminal control developments of those individuals cited in the references and to coworkers at Autonetics. Particular thanks are extended to J. W. Montooth, who performed some of the equation developments reported in the paper.

A Parameter-Perturbation Adaptive Control System*

R. J. McGRATH†, MEMBER, IRE, V. RAJARAMAN‡, STUDENT MEMBER, IRE, AND V. C. RIDEOUT‡, FELLOW, IRE

Summary—Theoretical and simulation studies of a parameter-perturbation self-adaptive system are discussed. A number of system block diagrams are included, showing diverse applications of parameter-perturbation adaptive techniques. A particular study has been made using error signals based on an ideal model, and the concept of high-frequency perturbation of the model has been introduced.

A linearized mathematical model of the adaptive loop for the system-adaptive scheme has been obtained using time-varying system analysis. Experimental verification of the mathematical model has been obtained with an analog computer. It is shown that increased speed of adaptive loop response is possible with high-frequency perturbation of model parameters.

Simulation studies have also shown the feasibility of adaptive control with random system input, and random parameter disturbances.

I. Introduction

DESPITE the rapidly increasing amount of attention now being given to self-adaptive systems, little attention has been paid to one of the forms of such systems, that is, the parameter-perturbation scheme. This method, first mentioned by Draper and Li,[1] was dealt with at some length in a recent paper by McGrath and Rideout.[2] Other investigators[3,4] have also given this method some recent attention, particularly in the field of process control.[5]

The parameter-perturbation self-adaptive scheme is only one of a number of effective feedback schemes. It is a general one in that it can permit an optimum adjustment of one or more parameters (according to a chosen performance measure) in spite of changes in any parameters or in input statistics. Good progress has been made toward maximization of its speed of response and elimination of perturbation effects from system output.

II. Block Diagrams of Adaptive Systems

Fig. 1(a) shows the basic parameter-perturbation adaptive system.[2] The error signal, in this case a "model error," is used to form an instantaneous error measure,

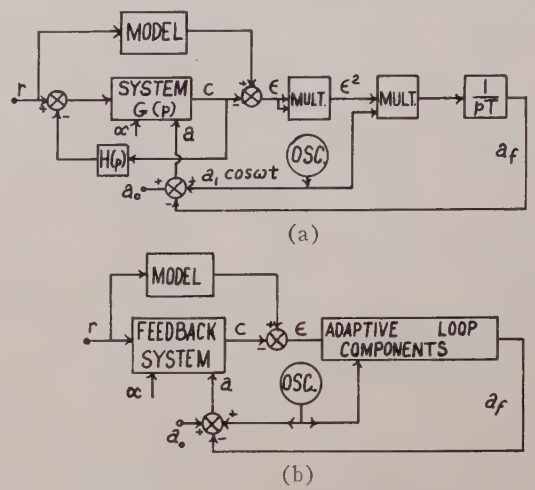


Fig. 1—Basic configuration of parameter-perturbation adaptive system.

which, here, is the error squared. This measure contains a component of the frequency ω at which a controllable parameter a is being perturbed. The amplitude and phase of this component give the magnitude and sign of a signal which can be recovered by multiplication and integration, as shown. This is fed back negatively to reduce the short time average of $\epsilon^2(t)$.

There may be a number of adaptive loops, each with a different oscillator frequency, and each controlling a different parameter. If the loops have the same gain, simultaneous control along a path of steepest descent will tend to result.[2] Also, additional elements in the loops may be desirable. These include a filter after the squarer, and a limiter after the multiplier.

As pointed out in a previous paper,[2] this system can be both signal-adaptive and system-adaptive. However, if the model is one to which the system may become identical under the influence of its adaptive loops, then it can only be system- (or parameter-) adaptive, because the error disappears when identity is reached no matter what the signal may be.

Fig. 1(b) is a simplified representation of the basic system shown in the first diagram, with only one adaptive loop shown, as before.

One difficulty encountered in the parameter-perturbation system is that the perturbation appears in the output. This may be avoided where a model is used which is of the same form as the system, or nearly so, by perturbing the parameter in the model corresponding to the parameter to be controlled in the system, as shown in Fig. 2. This has the further advantage, with high-frequency perturbation, of phase-shift problems being reduced. However, this scheme is limited to the case

* Received by the PGAC, December 15, 1960; revised manuscript received, March 15, 1961.

† Aerospace Corp., Los Angeles, Calif. Formerly with the University of Wisconsin, Madison, Wis.

‡ Dept. of Elec. Engrg., University of Wisconsin, Madison, Wis.

[1] C. S. Draper and Y. J. Li, "Principles of Optimalizing Control Systems and an Application to an Internal Combustion Engine," ASME Publications; September, 1951.

[2] R. J. McGrath and V. C. Rideout, "A simulator study of a two-parameter adaptive system," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-6, pp. 35–42; February, 1961.

[3] G. Vasu, "Experiments with optimalizing controls applied to rapid control of engine pressure with high amplitude noise signal," *Trans. ASME*, vol. 79, pp. 481–488; April, 1957.

[4] P. Eykhoff and O. J. M. Smith, "Optimalizing control with process dynamics identification," to be published in IRE TRANS. ON AUTOMATIC CONTROL.

[5] G. E. P. Box, "Some general considerations in process optimatization," *Trans. ASME*, JBE, vol. 82, pp. 113–119; March, 1960.

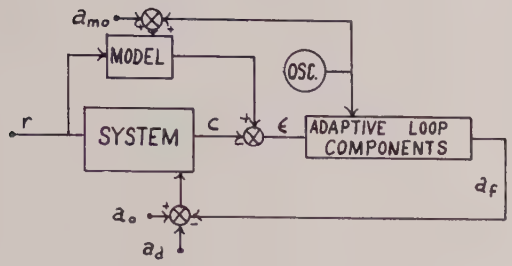


Fig. 2—A system-adaptive parameter-perturbation system.

where the disturbed parameter and the controlled parameter are essentially the same.

It is possible to use different error measures (ϵ^2, $|\epsilon|$, $|\epsilon|^n$, etc.) for the adaptive loops where more than one parameter is to be controlled. Also, different models may be used in forming the error. Fig. 3 shows one example of the use of multiple models. Here, the upper model and the adaptive loop are as in Fig. 2, with perturbation of this model, which must therefore be nearly identical to the system. The other adaptive loop uses perturbation of the system. Its model might be different from that of the system, so that changes in input statistics are emphasized in ϵ_2. Thus, in effect, the first loop corrects a to compensate for system changes α, and the second loop corrects b to compensate for input changes. Note that correction in b must be applied to model 1.

Another configuration which is both system-adaptive and signal-adaptive is shown in Fig. 4. Here, model 1 is optimized for changes in input statistics by adaptive loop 2 which operates on the error ϵ_2 between model 1, which is of the same form as the system, and model 2, which is of a different form such as to emphasize changes in input statistics. Adaptive loop 1 causes the system parameter a to follow the model parameter a_m which has been optimized for input statistics, and at the same time corrects for disturbances a_d.

The parameter-perturbation scheme may be used to make a plant "follower" as shown in Fig. 5. Here, a plant follower or the "learner" discussed by Margolis and Leondes[6] is caused by means of a parameter-perturbation adaptive loop (or loops) to follow changes in a plant which cannot otherwise be known and followed. Here the plant follower is a model, the parameters of which are forced to follow those of the plant according to the schemes discussed above. The information from the adaptive loops is then used to compute, according to predetermined formulas, desired changes in the controller parameters.

It has been mentioned that feed-forward should be used where possible.[7] Thus, information obtained by

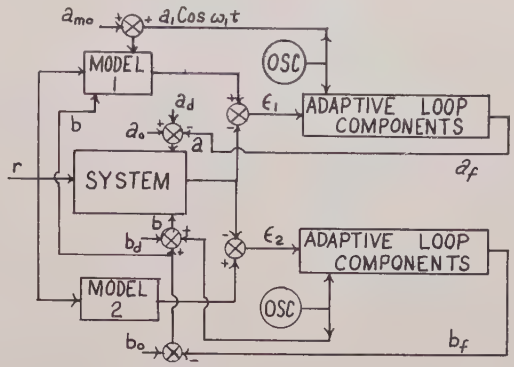


Fig. 3—A system and signal-adaptive scheme using multiple models.

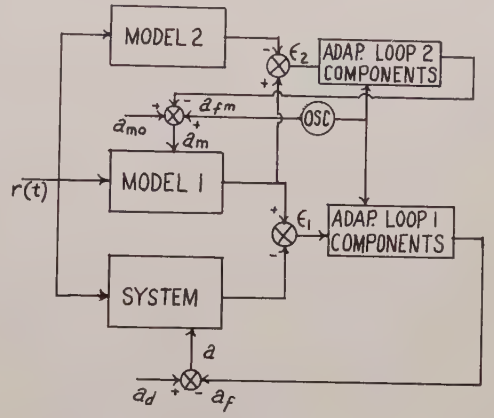


Fig. 4—A system and signal-adaptive scheme utilizing perturbation of model parameter.

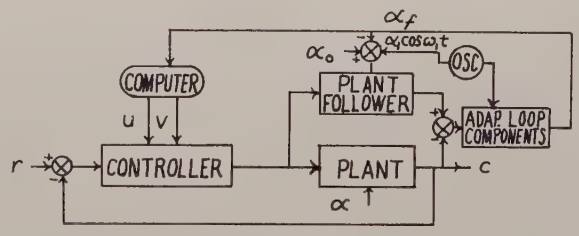


Fig. 5—Mechanization of Margolis-Leondes scheme using parameter perturbation.

processing the signal (by autocorrelation estimation,[8] for example), and by use of the outputs of various sensors which are related to system parameters, may be used as shown in Fig. 6 to control some parameters a and b of a system in a near-optimum fashion. An adaptive loop (or loops) may be added as also shown in Fig. 6 to "trim up" a parameter (or parameters), whether or not they are also to be controlled by feed-forward means.

The complexity of a parameter-perturbation loop is apt to be mentally perturbing. The circuit of Fig. 7 shows how such loops might be simplified. Here, the performance measure is square law (if the crystal rectifiers are square law), and filtering is accomplished in the

[6] M. Margolis and C. J. Leondes, "A parameter tracking servo for adaptive control systems," 1959 IRE WESCON Convention Record, pt. 4, pp. 104–115.

[7] J. Balchen, "The Use of Automatic Experimentation Combined with Mathematical Models in Optimalizing Control of Continuous Processes," presented at the Fifth Internatl. Instruments and Measurements Conf., Stockholm, Sweden; September, 1960.

[8] G. W. Anderson, R. N. Buland and G. R. Cooper, "The aeronutronic self-optimizing control system," *Proc. WADC Self-Adaptive Flight Controls Symp.*, Wright Air Dev. Ctr., Dayton, Ohio, Tech. Rept. 59–49; March, 1959.

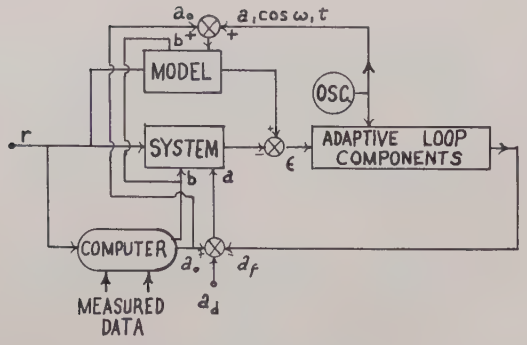


Fig. 6—An adaptive scheme using feed-forward and feedback control.

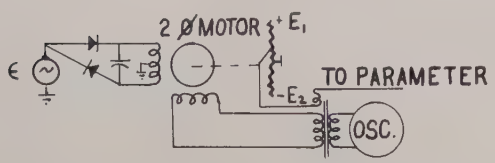


Fig. 7—A simple mechanization of the adaptive-loop components of parameter-perturbation schemes.

control phase of a two-phase motor by adding a shunt capacitor. The motor has its excitation phase fed by an oscillator (which may require but a single transistor). The motor acts as multiplier and integrator, the output being obtained from a shaft-driven potentiometer, as shown.

The detailed block diagram of the system discussed in the simulation studies is shown in Fig. 8. Suppose the system damping varies unpredictably over a large range because of changes in a. Also, let us assume that the controller gain b can be adjusted to compensate for this. The purpose of the adaptive loop is to identify variations in the parameter a, and suitably adjust b, so that ideally $a+b=2\zeta\omega_n$, the fixed model damping. (Here, ω_n is assumed invariant.)

In order to do this, ζ, in the model, is sinusoidally perturbed at a high frequency ν. Since the model and system are assumed identical, the only terms appearing in the error are due to the disturbances of the system (away from ideal), and the perturbation of the model. Assume a sinusoidal input of frequency η as the input to the system. (This particular input is considered for simplicity of explanation.) Let the power spectrum of disturbance be rectangular of width 2λ centered on zero frequency (assume $\nu \gg \eta \gg \lambda$). Fig. 9 shows the spectra of the signals at points (a), (b), (c) and (d) of Fig. 8 (the adaptive loop is assumed open). It is clear from Fig. 9(d) that at the point (d) in Fig. 8, the disturbance spectrum has been recovered, so that the information to control b is available.

It is apparent that the disturbance is identified by the combined effects of the input, perturbation and the adaptive loop components. Use of a cosine input has simplified the explanation. However, the system is capable of working with normal system inputs, except that

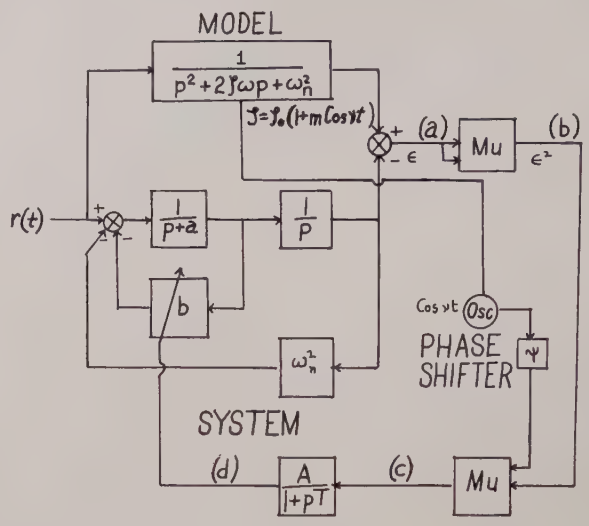


Fig. 8—Adaptive system used in simulation studies ($T \gg 1$).

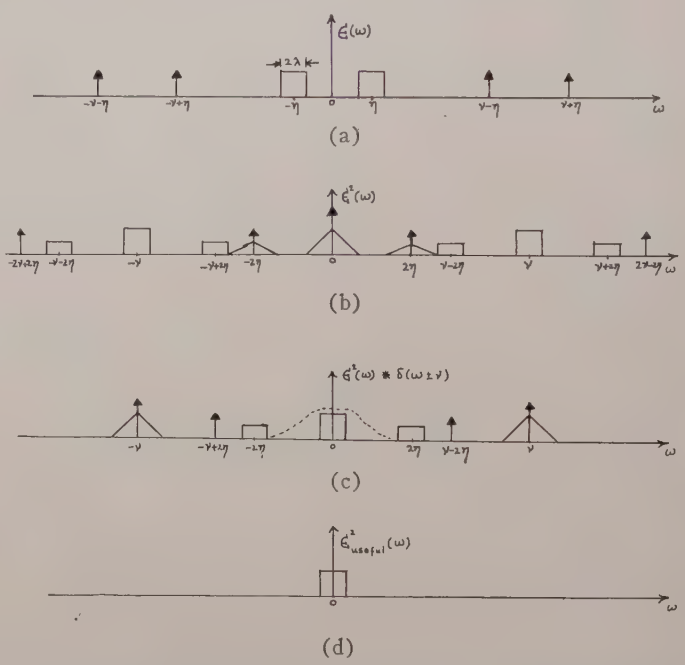


Fig. 9—Spectra of signals at points (a), (b), (c) and (d) of the system of Fig. 8 with adaptive loop open. Input to the system is assumed sinusoidal of frequency η. Disturbance spectrum assumed low-pass of width 2λ. Dotted line in (c) indicates the low-pass filter characteristics, and (d) shows the spectrum detected by adaptive loop. Higher-frequency components are not shown in (b) and (c).

the filtering problem is more stringent, and consequently the adjustment is slower. In fact, in the simulation, the input was a random process and the system was disturbed by an independent narrow-band low-frequency noise source. Under these realistic conditions, the adaptive scheme worked quite well.

It is of interest to note (Fig. 9) that the use of high-frequency perturbation has led to good filtering possibilities and consequently to rapid identification of disturbances. Thus, the speed of response of the adaptive loop is improved considerably by using a high-frequency perturbation of the model.

III. A Linearized Mathematical Model of the Adaptive Loop

Before deriving the linear model, a short review of time-varying systems is in order. In the time domain, a system output may be obtained from its input $x(t)$, by using

$$y(t) = \int_{-\infty}^{t} h(t, \tau)x(\tau)d\tau, \tag{1}$$

where $h(t, \tau)$ is the impulse response of the system. For time-invariant systems, this reduces to

$$y(t) = \int_{-\infty}^{t} h(t - \tau)x(\tau)d\tau. \tag{2}$$

If the input is $e^{j\omega t}$, then

$$y(t) = H(j\omega)e^{j\omega t}, \tag{3}$$

where $H(j\omega)$ is the Fourier transform of $h(t)$ and is called the transfer function.

If the system is time-varying with an input $e^{j\omega t}$, then

$$\begin{aligned} y(t) &= e^{j\omega t}\int_{0}^{\infty} h(t, t - \tau)e^{-j\omega\tau}d\tau \\ &= e^{j\omega t}H(j\omega; t) \end{aligned} \tag{4}$$

where $H(j\omega; t)$ is defined as the time-varying system function.

Consider the differential equation of a time-varying first-order system,

$$\frac{dy}{dt} + a(t)y(t) = x(t). \tag{5}$$

If $x(t) = e^{j\omega t}$, then using (4) we have

$$\frac{d}{dt}[e^{j\omega t}H(j\omega; t)] + a(t)H(j\omega; t)e^{j\omega t} = e^{j\omega t}$$

or

$$\frac{d}{dt}[H(j\omega; t)] + H(j\omega; t)[a(t) + j\omega] = 1. \tag{6}$$

The solution of (6) gives us the time-varying system function and allows us to solve for the response to other types of inputs; for, if we have an input $e_1(t)$ with Fourier transform $E_1(j\omega)$, then the output is given by

$$e_0(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(j\omega; t)E_1(j\omega)e^{j\omega t}d\omega. \tag{7}$$

Eq. (7) can often be evaluated using contour integration. This frequency-domain approach to time-varying systems is due to Zadeh.[9]

[9] L. Zadeh, "Frequency analysis of variable networks," Proc. IRE, vol. 38, pp. 291–299; March, 1950.

Consider the simple first-order feedback system with the transfer characteristic

$$C_a(j\omega) = \frac{1}{j\omega + b + a}, \tag{8}$$

and a model described by $1/(j\omega + M)$ (Fig. 10). Ideally, $(b+a)$ must equal M, which is the reference value. In the actual system, the parameter b changes unpredictably. We assume that b cannot be directly measured, nor can its variations be quickly determined from other data. The adaptive loop is used to control a so that $b+a$ approximately equals M at all times. The parameter disturbance away from the normal is detected with the aid of sinusoidal perturbation of M in the model.

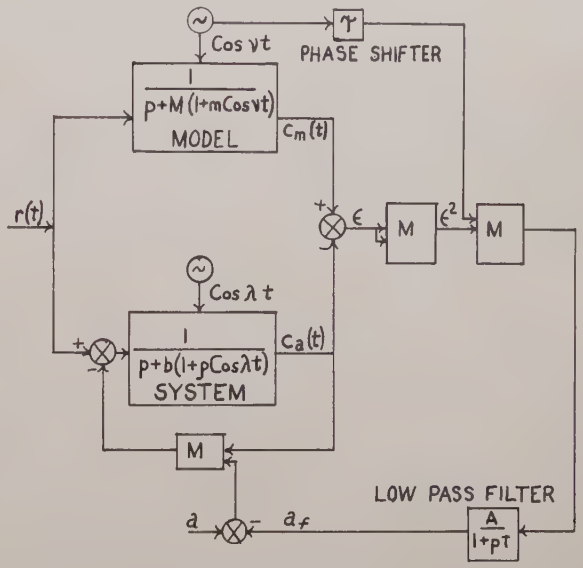


Fig. 10—Adaptive system used in the analysis. Here, ν is the frequency of perturbation of parameter M; λ is the frequency of disturbance of parameter b; a is the normal gain of controller, and p is the operator d/dt.

Consider the differential equation of the sinusoidally perturbed model:

$$\frac{dc_m}{dt} + M(1 + m\cos\nu t)c_m = r(t). \tag{9}$$

Here, m is the fractional change in M, and ν is the perturbation frequency. Other quantities are defined in Fig. 10. In order to find the time-varying system function corresponding to the above differential equation, we use the technique developed in deriving (6) and obtain

$$\begin{aligned} \frac{d}{dt}C_m(j\omega; t) + (M + j\omega)C_m(j\omega; t) \\ = 1 - Mm\cos\nu t C_m(j\omega; t), \end{aligned} \tag{10}$$

where $C_m(j\omega; t)$ is the time-varying system function of the model. The perturbation technique (proposed by Zadeh[9]) for solving this equation consists of first neglect-

ing the term with the time-varying coefficient and obtaining a solution $C_{m0}(j\omega;\ t)$. Successive solutions are obtained from

$$\frac{d}{dt}\,C_{mn}(j\omega;\ t) + (M + j\omega)C_{mn}(j\omega;\ t) = -\,Mm\cos\nu t C_{m(n-1)}(j\omega;\ t). \quad (11)$$

The complete solution is then

$$C_m(j\omega;\ t) = \sum_{n=0}^{\infty} C_{mn}(j\omega;\ t). \quad (12)$$

This summation develops into a Fourier series with fundamental frequency ν, with the coefficients as a power series in $-mM/(M+j\omega+j\nu)$. Since

$$\left|\frac{mM}{M + j\omega + j\nu}\right| < m \ll 1,$$

in the approximate analysis we retain only the first two terms. Thus, we have

$$C_m(j\omega;\ t) \approx \frac{1}{M + j\omega} + \left[\frac{-mM}{2}\,\frac{1}{M + j\omega}\right] \cdot \left[\frac{e^{j\nu t}}{M + j\omega + j\nu} + \frac{e^{-j\nu t}}{M + j\omega - j\nu}\right]. \quad (13)$$

Our aim is to obtain a linear equivalent of the open adaptive loop which detects changes in b. Therefore, we assume that the variation of b is sinusoidal and is given by $b(1+\rho\cos\lambda t)$, where ρ is the per unit variation of b from normal, and λ the frequency of disturbance. The differential equation of the sinusoidally-disturbed system is

$$\frac{dc_a}{dt} + b(1 + \rho\cos\lambda t)c_a + ac_a = r(t). \quad (14)$$

We note the similarity of (9) and (14), and following the same reasoning, obtain for the approximate system function of the disturbed system:

$$C_a(j\omega;\ t) \approx \frac{1}{M + j\omega} + \left[\frac{-b\rho}{2}\quad\frac{1}{M + j\omega}\right] \cdot \left[\frac{e^{j\lambda t}}{M + j\omega + j\lambda} + \frac{e^{-j\lambda t}}{M + j\omega - j\lambda}\right]. \quad (15)$$

The error is formed by taking the difference between the model and the system outputs. Thus, using (13) and (15), we obtain for the transform of error

$$E(j\omega;\ t) = \frac{1}{M + j\omega}\left[\left(\frac{-mM}{2}\right)\left(\frac{e^{j\nu t}}{M + j\omega + j\nu} + \frac{e^{-j\nu t}}{M + j\omega - j\nu}\right) - \left(\frac{-b\rho}{2}\right)\left(\frac{e^{j\lambda t}}{M + j\omega + j\lambda} + \frac{e^{-j\lambda t}}{M + j\omega - j\lambda}\right)\right]. \quad (16)$$

We now assume that the input to the model-system combination is a fixed amplitude sinusoid given by

$$r(t) = R\cos\eta t. \quad (17)$$

With this input, the "time-domain" expression of the error [obtained by inverse transforming of (16)] is given by

$$\epsilon(t) = \frac{R}{2(M^2 + \eta^2)^{1/2}} \cdot \left[\frac{(-Mm)}{\{M^2 + (\eta + \nu)^2\}^{1/2}}\cos\left(\overline{\nu + \eta}\,t - \phi_1 - \phi_2\right) + \frac{(-Mm)}{\{M^2 + (\eta - \nu)^2\}^{1/2}}\cos\left(\overline{\nu - \eta}\,t + \phi_1 - \phi_3\right) - \frac{(-b\rho)}{\{M^2 + (\eta + \lambda)^2\}^{1/2}}\cos\left(\overline{\eta + \lambda}\,t - \phi_1 - \phi_4\right) - \frac{(-b\rho)}{\{M^2 + (\eta - \lambda)^2\}^{1/2}}\cos\left(\overline{\eta - \lambda}\,t - \phi_1 + \phi_5\right)\right],$$

where

$$\phi_1 = \tan^{-1}\frac{\eta}{M},\quad \phi_2 = \tan^{-1}\frac{\eta + \nu}{M},\quad \phi_3 = \tan^{-1}\frac{\nu - \eta}{M},$$

$$\phi_4 = \tan^{-1}\frac{\lambda + \eta}{M},\quad \text{and}\quad \phi_5 = \tan^{-1}\frac{\lambda - \eta}{M}. \quad (18)$$

If we now assume that $\nu \gg \eta \gg \lambda$, (18) is greatly simplified. This assumption implies that: 1) we perturb the model parameter at a high frequency; 2) the disturbance of the system parameter is much slower than perturbation frequency; 3) the input frequency is between these two frequencies; and 4) the disturbance frequency is much smaller than the cutoff frequency of the system. We are solely interested in the "detected" signal at the output of the low-pass filter (Fig. 10). Thus, after squaring $\epsilon(t)$, we retain only the terms useful for this purpose and they reduce to a single term given by

$$\epsilon^2_{\text{useful}}(t) = \frac{1}{2}\cdot\frac{R^2 bM_m\rho\cos\phi_1}{(M^2 + \eta^2)^{3/2}(M^2 + \nu^2)^{1/2}} \cdot [\cos(\nu t - \phi_2)\cos\lambda t], \quad (19)$$

where $\phi_1 = \tan^{-1}\eta/M$ and $\phi_2 = \tan^{-1}\nu/M$. This signal is multiplied by $\cos(\nu t - \psi)$. It is then filtered by a narrow-band low-pass filter intended to attenuate frequencies much greater than λ. If we suppose that the low-pass filter has a transfer function $(A/(1+pT)$, we may represent the scheme used to detect the parameter disturbance $b\rho\cos\lambda t$ by the linear model shown in Fig. 11. If the disturbance, instead of being a sinusoid, is a low-frequency narrow-band signal $\Delta b(t)$, the same model is a good approximation of the "detection scheme" used in the adaptive servo.

In the adaptive servo, the detected parameter disturbance is fed back negatively to the controllable parameter. This provides a closed-loop control of the parameter. Using the linear model of Fig. 11, we may represent

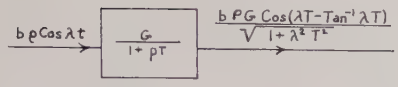


Fig. 11—Open-loop linear equivalent representing detection of parameter disturbance.

$$G = \frac{R^2 M m A \cos\phi_1 \cos(\psi - \phi_2)}{4(M^2 + \nu^2)^{1/2}(M^2 + \eta^2)^{3/2}}.$$

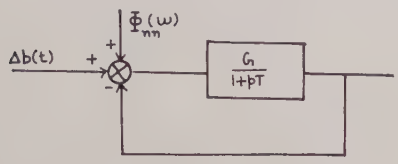


Fig. 12—Mathematical model of parameter servo.

this closed-loop parameter control by the "parameter servo" of Fig. 12. This mathematical model provides insight into the operation of the original adaptive system. In this model, $\Delta b(t)$ is the incremental time variation of the uncontrollable parameter. The function $\Phi_{nn}(\omega)$ represents "noise spectrum," which includes spectral components neglected in the analysis and also the noise in the system.

We observe that the gain of the parameter servo is proportional to the amplitude of the perturbation m, the square of the input amplitude R^2, the system parameter, the frequency of the input and the perturbation frequency. The time constant in the open-loop case is solely dependent on the low-pass filter. Thus, any change in the input amplitude and frequency will affect only the open-loop gain of the parameter servo.

The above analysis assumes a constant amplitude sinusoidal signal as the input to the system. If it is any other signal, the analysis is more complicated, but the model of the parameter servo will still be a first-order system with a different open-loop gain.[10] The noise spectrum $\Phi_{nn}(\omega)$ will also be different.

If the actual system and model are of higher order and only one parameter is controlled, the analysis is essentially the same with some minor modifications.[10]

It is important to remember that the mathematical model is true only as long as the assumptions made in deriving it are valid. In particular, the assumption that the disturbance is at a much smaller frequency compared to the cutoff frequency of the system simplified the dynamic model of the adaptive loop. In effect, the phase shift introduced at the disturbance frequencies by the system is much smaller than that introduced by the low-pass filter in the adaptive loop, and was neglected. For higher disturbance frequencies, this assumption is not valid. Thus, we cannot expect the simplified mathematical model to predict stability of the over-all system under all possible conditions.

It should also be mentioned that for the simple case considered with sinusoidal input, it is not essential to use the powerful techniques of Zadeh. The time-varying differential equations can be solved directly using perturbation techniques, leading to the results in this section. However, it is necessary to use the concept of time-varying system functions if the analysis is to be extended to the case with random inputs.[10]

[10] V. Rajaraman, "Theory of Parameter-Perturbation Adaptive and Optimizing Control Systems," Ph.D. dissertation, University of Wisconsin, Madison, to be published.

IV. Experimental Verification of the Proposed Linear Equivalent of the Parameter Servo

In order to check the validity of the proposed linear equivalent for the adaptive loop, simulator studies were conducted. The system of Fig. 10 was used with $M = 0.5$, $a = 0.4$, $b = 0.1$, $m = 0.1$ and the input was a sinusoid $r(t) = R \cos \eta t$. As shown in the figure, b was disturbed sinusoidally at a frequency λ rad/sec. The parameter M in the model was perturbed at frequency ν. In the computer used, the computer unit of time was 1 msec, so that the actual frequencies used, as indicated below, are 1000 times greater than if real time simulation had been used.

The first experiment was to check the linearity of the adaptive loop in detecting the disturbance. The values of η and ν chosen are 50 cps and 400 cps, respectively. R was held constant at 12.5 volts peak-to-peak. With λ fixed at 1 cps, ρ was varied. The plot of the output of the adaptive loop vs ρ [(a) when the loop is open, and (b) when it is closed] are shown in Fig. 13. This demonstrates the validity of the assumption of linearity made in the analysis.

The next experiment was to check the equivalent form of Fig. 12. In this case, the value of ρ was fixed at 0.5. The frequency of the disturbance λ was varied. The ampitude and phase of the disturbance signal at frequency λ detected at the output of the adaptive loop (with the loop open) was measured. The effective gain of the adaptive loop is the ratio of the amplitude of the detected disturbance to the amplitude of the input disturbance. The phase shift between the input and detected disturbances was also measured. The individual points in Fig. 14 correspond to these measured quantities after normalization. The continuous curves correspond to the normalized gain and phase plots of the open-loop parameter servo proposed in Section III. The gain and phase for the closed-loop parameter servo are also plotted as continuous curves in the same figure. The experimentally observed values (normalized) with the adaptive loop closed are the individual points. In both the open- and closed-loop cases, it is seen that the theoretical analysis predicts closely the experimental observations.

V. Measurement of Speed of Response of an Adaptive System

The measurement of the speed of response of an adaptive system requires that the signals and disturbances, as well as system parameters, be specified. In addition, a performance measure must be adopted if comparison of different systems and/or optimization of the adaptive loop is to be attempted.

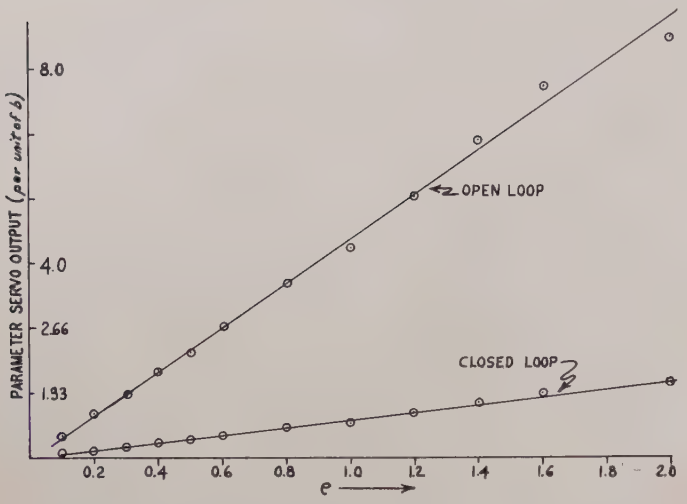


Fig. 13—Check on the linearity of parameter servo.

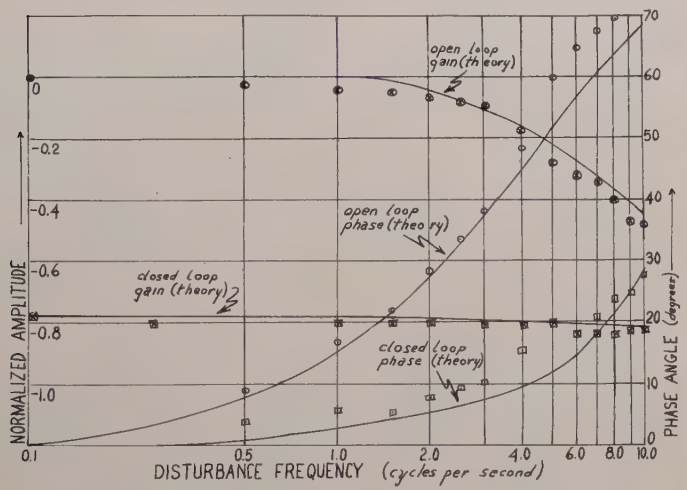


Fig. 14—Amplitude and phase characteristics of parameter servo. Individual points represent experimentally observed values. Legend: ⊗ open-loop gain; ⊙ open-loop phase; ⊠ closed-loop gain; and ⊡ closed-loop phase.

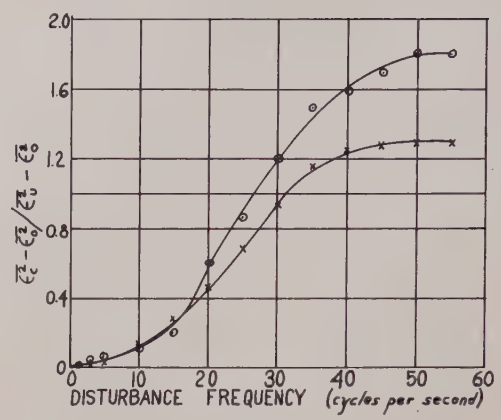


Fig. 15—Curves connecting the ratio of mean square errors and parameter disturbance, for 100-cps sinusoidal input signal, and for two values of ρ, the per unit peak change in parameter value. $\overline{\epsilon_0^2}$ is the mean squared error (MSE) with no disturbance, $\overline{\epsilon_c^2}$ the MSE with the adaptive loop closed and $\overline{\epsilon_u^2}$ the MSE with the adaptive loop open.

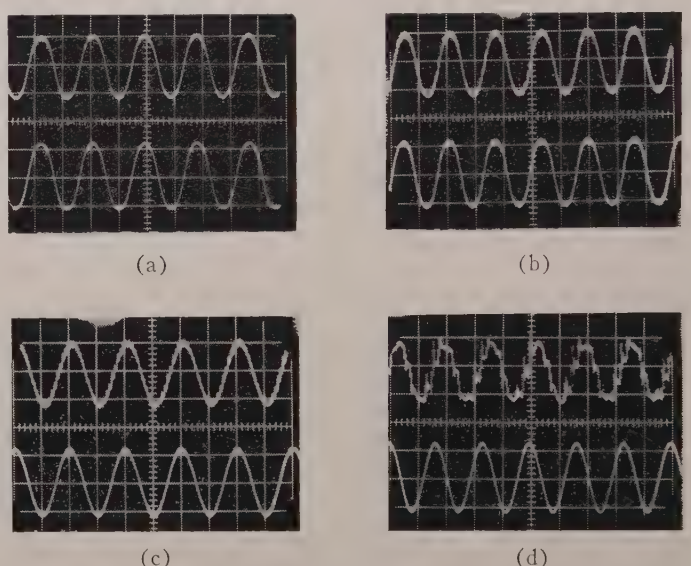


Fig. 16—Photographs showing the disturbance (lower traces in each case) and correction signals (upper trace) fed to the controller (obtained with adaptive loop closed). Disturbance frequencies are (a) 1 cps, (b) 5 cps, (c) 10 cps and (d) 30 cps. Input to the system in this case is a sine wave of 100 cps.

The system chosen for simulator studies of response speeds is shown in Fig. 8. The characteristic frequency of the second-order system was 169 cps (or 1000 rad/sec), and the input signal first used was a 100-cps sinusoid. The damping ratio of the model, set at 0.5, was sinusoidally perturbed over a large range (2.6 to 1) by a signal at 1150 cps. Because of the relatively high frequency of the perturbation, the model output at perturbation frequency was very small. The parameter a of the system was sinusoidally disturbed, as given by the equation $a = a_0(1 + \rho \cos \lambda t)$. Various values of ρ and λ were used to study adaptive speeds of response. Mean square error (MSE) between the system and model outputs, or a short-time approximation thereto, was used as a performance measure.

Fig. 15 shows graphs of measured relative MSE of the system as a function of disturbance frequency for two disturbance amplitudes. The adaptive controller in this case is able to substantially decrease the MSE for frequencies up to 20 cps, which is about an eighth of the system characteristic frequency. Fig. 16 is a series of photographs showing the disturbance (lower trace in each case) and the negative of the correction signal for $\rho = 0.25$. In Fig. 16, the disturbance (a) is at $\lambda/2\pi = 1$ cps, (b) 5 cps, (c) 10 cps, and (d) 30 cps. For the 30-cps case, the adaptive control signal has a noticeable component at the perturbation frequency, but these high frequencies are damped out by the system and do not appear in the system output.

The curves shown in Fig. 15 are for an optimum adaptive loop-gain setting for the input level used, and for very low disturbance frequencies. Improved results at higher disturbance frequencies are possible by appropriate loop-gain adjustment. Also, because adaptive loop gain varies with input signal level, the results obtained will differ if input level changes, unless the adaptive loop gain is suitably changed. This points up the necessity of AGC on adaptive loop, or use of limiters

before the integrator.[2] Effects of input level changes may also be reduced by choice of a suitable error criterion.[11]

It is also of very great interest to note that under a slightly different set of conditions with the system operating with a 50-cps input signal, it was found possible to have a MSE ratio of 0.5 with a 130-cps disturbance signal, that is, a disturbance of frequency *above* the input signal, and very close to system cutoff.

As an additional test, white Gaussian noise, shaped by a second-order filter tuned to 85 cps and with a damping ratio of 0.5, was used as a system input. Fig. 17(a) and (b) shows curves for MSE ratio for sinusoidal disturbances of two different amplitudes with this random input signal. Results are not as good as for sinusoidal input (Fig. 15) at low disturbance frequencies because of the low-frequency random noise which appears in the adaptive loop. At higher disturbance frequencies, the larger error caused by failure of the adaptive loop to follow these higher frequencies masks the low-frequency noise and results are more comparable with those of Fig. 15.

Fig. 18 shows the parameter disturbance and signal feedback to the controller when the input to the system consists of a random signal.

Fig. 19(a) and (b) shows the response to a step disturbance for the sinusoidal and random input cases, respectively. It should be noted that in keeping with the analytical results of Section III, the step response tends to be that of a first-order system.

Finally, a random parameter disturbance was used. An independent noise source was shaped to give a band of disturbance frequencies centered at 3-cps. Fig. 20(a) and (b) shows samples of the disturbances and the resultant correction signals for (a) sinusoidal and (b)

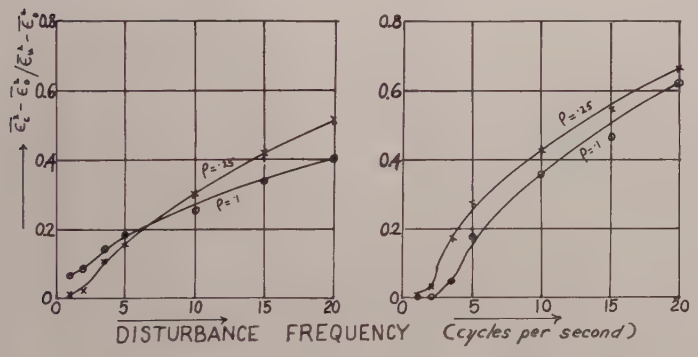


Fig. 17—Curves connecting the ratio of mean square errors and parameter disturbance. The input to the system is a Gaussian random process of mean square value (a) 1.2 and (b) 0.6 units.

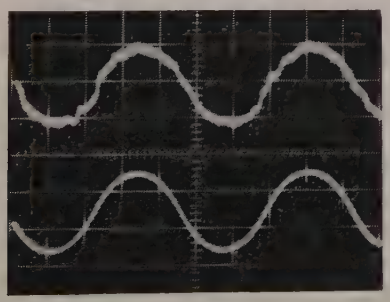


Fig. 18—Sinusoidal parameter disturbance (lower curve) at 1 cps and signal feedback to the controller (upper curve) with a random input signal to the system.

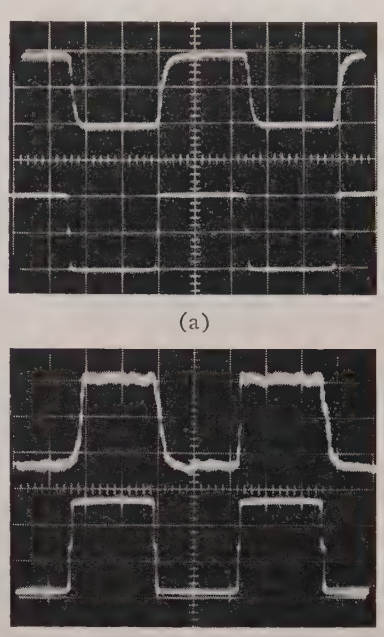


Fig. 19—Square-wave parameter disturbance, at 1 cps and signal fed to controller (upper curve); (a) with 100-cps sinusoidal input to system and (b) with random input to system.

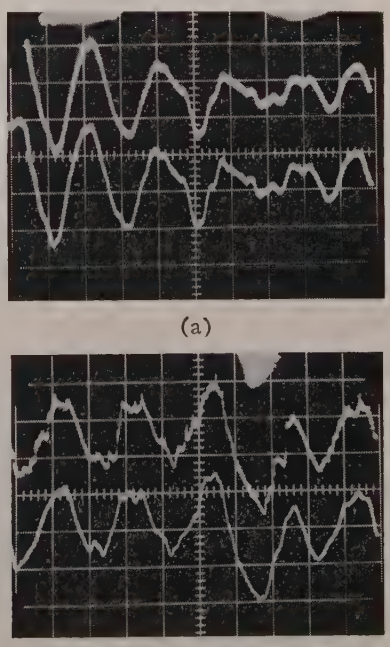


Fig. 20—Random parameter disturbance (lower curve) and signal feedback to the controller (upper curve) for (a) sinusoidal input to system and (b) random input to system. Time scale: 1 division = 200 msec.

[11] W. C. Schultz and V. C. Rideout, "Control systems performance measures; past, present, and future," IRE Trans. on Automatic Control, vol. AC-6, pp. 22–35; February, 1961.

random system inputs. It is particularly encouraging to note that the system works well when both the input signal and parameter disturbance are random processes.

VI. Conclusions

The parameter-perturbation scheme of adaptive control discussed in an earlier paper has been extended. These extensions are the use of model errors, perturbation of the model,[12] and combination with various feedforward, and other, schemes. The use of high-frequency perturbation of the model parameter has been shown to give an increased speed of adaptation for the system-adaptive scheme.[10] Perturbing a model is more practical, and has the further advantage of eliminating the appearance of the perturbation at the output of the actual system.

A mathematical model has been obtained using the techniques of time-varying system analysis. This model is capable of rather general application to parameter-perturbation adaptive schemes. The model which turns out to be of simple form provides insight into the operation of the adaptive loop, and indicates the important aspects in its design. Further work along these lines are directed at extending the model to cases where: 1) the system input is a random process, and 2) the system is signal-adaptive (Fig. 4).

Simulation studies have provided checks on the linearized model of the adaptive loop, and have provided further information on the behavior of the adaptive system with random inputs. In particular, it has been shown that this system can correct for parameter disturbances of frequencies as high as one eighth of the system cutoff frequency.

Further theoretical and simulator studies are being conducted on an optimalizing chemical process which uses parameter-perturbation techniques. This scheme was proposed independently by Box.[5] It is of interest to note that the mathematical model of the adaptive loop in this case too, has the same form as the one proposed in this paper.[10]

[12] P. Eykhoff, "Optimizing Control and Process Parameter Estimation," Doctoral dissertation, Univ. of California, Berkeley; 1960.

Transfer-Function Tracking and Adaptive Control Systems*

C. N. WEYGANDT† and N. N. PURI‡

Summary—**This paper describes a system for determining the parameters of a transfer function of the form: a constant divided by a polynomial in the Laplacian variable *s*. The system is described in detail for a polynomial of second order, but may be extended to polynomials of any order. The system may be realized in terms of ordinary analog computer components. It presents the information in a form which may be readily inserted into a controller for realization of a general type of adaptive control system.**

I. Introduction

THE word "adaptive" has been appearing in the title of many papers [1]–[3], dealing with the general field of control in the past few years. It has almost as many shades of meaning as the number of papers in which it appears. For the present purposes, let us define an adaptive control system as follows:

Suppose there exists a controlled system where:

1) The performance of the system is describable by a system of equations, but the parameters of the equations, and perhaps even their form, are not known in detail.
2) The system is affected in some manner by the environment in which it resides.

Then, the system is adaptive if it is capable of:

1) Determining its performance equations and the effect of its environment in sufficient detail so that:
2) Some performance criterion shall be always optimum in spite of arbitrary variations in the effect of the environment.

In particular, we shall confine ourselves to the case where the controlled system is describable by a set of linear differential equations with coefficients which, while they may vary with time because of the effect of environment, vary sufficiently slowly so that they may be considered constant for the purposes of calculating

* Received by the PGAC, November 15, 1960; revised manuscript received, March 9, 1961. This paper is a condensation of part of a Ph.D. dissertation at the University of Pennsylvania, Philadelphia, 1961.

† Moore School of Elec. Engrg., University of Pennsylvania, Philadelphia, Pa.
‡ Dept. of Elec. Engrg., Drexel Inst. Tech., Philadelphia, Pa.

the performance criterion at any time. The order of the system of differential equations need not be known.

The environment may affect the system, for example, through the input, the output or load, or through variations in the coefficients. The dynamic behavior of the kind of system considered here may be described by a differential equation of the form

$$\sum_{k=1}^{r} a_k \frac{d^k}{dt^k} x = g(x), \tag{1}$$

where the a_k describe the system, x is the system variable, and $g(x)$ is any function, linear or nonlinear, of x.

To control the system, we apply a force f. The equation of the controlled system then becomes

$$\sum_{k=1}^{r} a_k \frac{d^k}{dt^k} x = g(x, f). \tag{2}$$

If we restrict ourselves to linear systems, (2) becomes

$$\sum_{k=1}^{r} a^k \frac{d^k}{dt^k} x = -a_0 x + f(x), \tag{3}$$

where $f(x)$ is a linear function of x.

Control is frequently effected by making $f(x)$ depend on the difference between the actual x and some desired value x_0

$$\sum_{k=1}^{n} c_k \frac{d^k}{dt^k}(x - x_0) = \sum_{k=1}^{m} d_k \frac{d^k}{dt^k} f(x), \tag{4}$$

where the c's and d's describe the controller. For a physically realizable controller, $m \geqq n$.

Taking Laplace transforms of (3) and (4) with all initial conditions equal to zero, we get

$$\sum_{k=0}^{r} a_k s^k X = F = \frac{\sum_{k=1}^{m} c_k s^k}{\sum_{k=1}^{n} d_k s^k}(X - X_0). \tag{5}$$

Capital letters represent Laplace transforms of corresponding lower case letters. Eq. (5) becomes

$$H^{-1}X = B(X - X_0),$$

where

$$H^{-1} = \sum_{k=1}^{r} a_k s^k$$

$$B = \frac{\sum_{k=1}^{n} c_k s^k}{\sum_{k=1}^{m} d_k s^k}. \tag{6}$$

H is the transfer function of the system to be controlled, or the plant, and B is the transfer function of the controller.

II. A Particular System

Fig. 1 is a block diagram of a complete system to track the transfer function $1/s^2+as+b$. The part in the dashed lines is an ordinary feedback control system with input x_0 and output x. H is the system to be controlled or plant, and B is the controller. The open-loop transfer function is

$$\frac{x}{e} = BH = \frac{s^2 + \alpha s + \beta}{s^2 + cs + d} \times \frac{1}{s^2 + as + b}, \tag{7}$$

where a and b are particular values of the plant coefficients a_r. These are assumed to vary slowly with time in an unknown fashion. c and d are particular values of the controller coefficients d_m which are known constants. α and β are approximations to a and b. The purpose of the adaptive part of the system is to force α and β to be equal to a and b, respectively. This is what is meant by tracking.

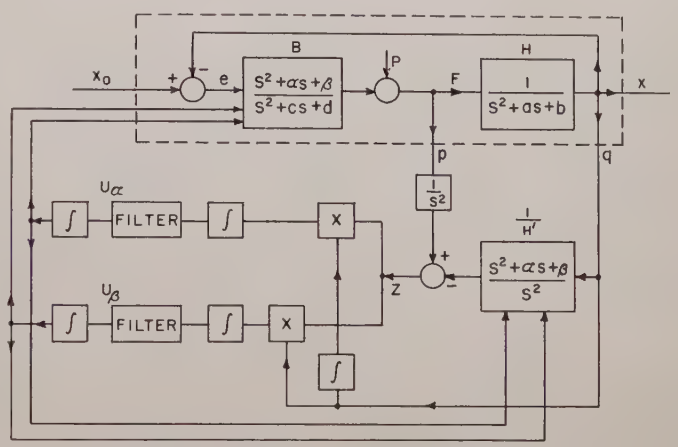


Fig. 1—Block diagram of the complete system.

The rest of the diagram, representing a system which will assure that $\alpha = a$ and $\beta = b$, will now be described and analyzed. The system requires signals from the input and output of the plant at p and q, and the application of a perturbing signal at P. Access to the controller is also necessary to make adjustments in α and β.

In general, the system operates as follows. The perturbing signal P is a sinusoid which excites H. The resulting signal x is fed into $1/H'$, and the output of $1/H'$ is compared with the double integral of the input to H, where H' is an approximate to H. The resulting signal Z can be shown (see Appendix II) to be

$$Z = (a - \alpha) \int M \sin(\omega t + \theta) dt + (b - \beta) \iint M \sin(\omega t + \theta) dt. \tag{8}$$

This signal contains information about the deviation of α from a, and β from b, combined in one signal. It is necessary to separate this information to accomplish

separate control of α and β. To do this, Z is multiplied by x, and the resulting product is integrated and filtered. The resulting signal u_β can be shown to be

$$u_\beta = (b - \beta)\frac{M^2}{2\omega^2} - \frac{\dot{u}_\beta}{\gamma}, \tag{9}$$

where γ is a parameter of the filter. The signal u_β is independent of α because of the orthogonality property of sines and cosines. That is, the average value of

$$\int \sin \omega t \cos \omega t dt = 0$$

$$\int \sin^2 \omega t dt = \tfrac{1}{2}. \tag{10}$$

Similarly, except that Z is multiplied by $\int x dt$, we generate the signal

$$u_\alpha = (a - \alpha)\frac{M^2}{2\omega^2} - \frac{\dot{u}_\alpha}{\gamma}. \tag{11}$$

This use of the orthogonality of the sine and cosine of a single frequency to extract separate information from a single signal for the purpose of parameter tracking is believed to be presented here for the first time.

These signals u_α and u_β, when then made equal to $\dot{\alpha}/K_\alpha$ and $\dot{\beta}/K_\beta$, respectively, give rise to the differential equations

$$\ddot{\alpha} + \gamma\dot{\alpha} + \alpha\frac{K\alpha M^2}{2\omega^2}\gamma = a\frac{K\alpha M^2}{2\omega^2}\gamma \tag{12}$$

and

$$\ddot{\beta} + \gamma\dot{\beta} + \beta\frac{K\alpha M^2}{2\omega^2}\gamma = b\frac{K\alpha M^2}{2\omega^2}\gamma. \tag{13}$$

These equations assure that in steady state, α will equal a, and β will equal b. The filter parameter γ and the tracking loop gains K_α and K_β can be chosen so that changes in α and β take place slowly compared to the responses of the plant, which justifies the assumption that α and β were constant in the first part of the analysis. Fortunately, this requires γ to be small, which is also the condition for adequate filtering of the multiplier output.

The frequency of the perturbing sinusoid must lie in the pass band of the plant transfer function, or there will not be enough variation in x to drive the tracking loops. If the plant has a natural frequency (or range of natural frequencies, since its parameters are assumed to vary), this range should be avoided. The perturbing signal is a necessary evil of this method. If we want to discover the dynamic characteristics of a system, we must excite it in some way so that its characteristics are made manifest. The perturbation need not be sinusoidal; perhaps there are other periodic functions that are known or could be invented that have the same kind of orthogonality properties.

In practice, the perturbing signal would not have to be present all the time—perhaps only for one minute out of each hour, or ten minutes out of each day, depending on how quickly the system parameters change with time.

To complete the description of the system, the signals

$$u_\alpha = \dot{\alpha}/K_\alpha$$

and

$$u_\beta = \dot{\beta}/K_\beta \tag{14}$$

are integrated, and the integrator outputs are used to adjust the values of α and β, both in the controller and in $1/H'$. Thus, whenever the perturbing signal is present, the system will be adjusted so that its dynamic performance is independent of the plant parameters.

If, in addition to the perturbation at P, there is a perturbation at x_0 or any other point in the system, it will in no way affect the corrective signals from the adaptive loops, unless the second perturbation happens to give rise to a frequency equal to the frequency ω of the perturbation at P. This is why ω should not lie in range of natural frequencies of the plant. The system described and a number of similar variatons have been simulated on an analog computer and found to operate very satisfactorily. During this simulation, step function perturbations were inserted at x_0, and it was found that the tracking loops were not materially affected.

A first look at Fig. 1 suggests that the tracking loops with their many integrations might create a stability problem. However, on analysis, the tracking loops give rise to (12) and (13), which cannot possibly be unstable. Even if higher derivatives of α and β were present, a proper choice of K_α and K_β would make the system stable. What actually happens is that the orthogonality properties of the sines and cosines make the adaptive loops essentially independent of the system to be controlled as far as stability is concerned.

III. Extensions to Higher-Order Systems and Multiloop Systems

The general method presented here can be extended to systems where the denominator is of any order. It is necessary to have $n/2$ different frequencies where n is the order of the system. This is an improvement by a factor of two over other methods which require a different frequency for each parameter tracked. This method, however, cannot be simply extended to systems where the numerator is not a constant. This is a serious limitation.

The order of the denominator need not be known. If it were assumed to be of sixth order, and six tracking loops were provided, but it turned out that the system was only of fifth order, the sixth parameter would turn out to be zero.

This general system can be extended to multiloop systems where the plant has more than one aspect to be controlled. This extension will be the subject of a later paper.

IV. Conclusions

A system of parameter tracking has been presented which:

1) Requires the minimum of perturbing signals.
2) Requires access to the controlled plant only at places where such access should be easy in practice.
3) Is capable of physical realization with standard, well-tried analog computer components.
4) Produces the values of the tracked parameters in such a form that they can be readily inserted into the controller to complete an adaptive control system.

Appendix I

Analysis of Tracking Loops

From Fig. 1 the signal F applied to the plant is

$$F = P - Be, \tag{15}$$

where

$$P = \frac{\omega}{s^2 + \omega^2} \text{ for sinusoidal perturbation,}$$

and

$$B = \frac{s^2 + \alpha s + \beta}{s^2 + cs + d}.$$

α and β are considered constant although they vary slowly.

$$F = \frac{\omega}{s^2 + \omega^2} - \frac{s^2 + \alpha s + \beta}{s^2 + cs + d}(x_0 - x). \tag{16}$$

If the system is a regulating system and we are considering incremental variations about a constant set point, $x_0 = 0$. Then

$$X = \frac{1}{s^2 + as + b} F \tag{17}$$

$$F = \frac{\omega}{s^2 + \omega^2} + \left(\frac{s^2 + \alpha s + \beta}{s^2 + as + b}\right)\left(\frac{F}{s^2 + cs + d}\right), \tag{18}$$

from which

$$X = M(s) \frac{\omega}{s^2 + \omega^2},$$

where $M(s)$ is a ratio of polynomials in s having a fourth-degree denominator. The time function corresponding to X,

$$X(t) = M \sin(\omega t + \theta) + \text{transient terms}, \tag{19}$$

where $M = |M(s)|_{s=j\omega}$ and θ is the angle of complex number resulting from the substituting of $j\omega$ for s in $M(s)$. The frequency ω is so chosen that the transient terms have essentially died to zero in one or two cycles of the sinusoid. M is really a variable since it contains α and β, but if the system is close to equilbrium, α is nearly a, and β is nearly b so that M will not vary much. Hereafter, we will consider that $X(t)$ is a pure sinusoidal of constant magnitude.

Now let us compute

$$Z(s) = \frac{F}{s^2}\left[1 - \frac{s^2 + \alpha s + \beta}{s^2 + as + b}\right]. \tag{20}$$

The s^2 is inserted in the denominator of $1/H'$ to make it readily realizable with analog computer components. In a more general case, the denominator of $1/H'$ would be s^n where n is the degree of the polynomial being tracked.

$$Z(s) = X(s)\left[\frac{a - \alpha}{s} + \frac{b - \beta}{s^2}\right], \tag{21}$$

and the time function corresponding to $Z(s)$ is

$$Z(t) = (a - \alpha)\int M \sin(\omega t + \theta)dt + (b - \beta)\iint M \sin(\omega t + \theta)dt,$$

or

$$Z(t) = -(a - \alpha)\frac{M}{\omega}\cos(\omega t + \theta) - (b - \beta)\frac{M}{\omega^2}\sin(\omega t + \theta). \tag{22}$$

This is (8) in the body of the paper.

The next step is to multiply $Z(t)$ by $X(t)$. In the frequency range of most control systems, this can be done with a servo-multiplier.

$$Z(t)X(t) = -(a - \alpha)\frac{M^2}{\omega}\cos(\omega t + \theta)\sin(\omega t + \theta) - (b - \beta)\frac{M^2}{\omega^2}\sin^2(\omega t + \theta). \tag{23}$$

This result is now integrated indefinitely. It would be nice to integrate over a definite integral number of periods of $\sin \omega t$, but no equipment is presently available for doing this.

$$\int ZXdt = -(b - \beta)\frac{M^2}{2\omega^2} + \text{terms in sine and cosine of } 2\omega t. \tag{24}$$

Similarly,

$$\int Z\left[\int X dt\right] dt = +(a-\alpha)\frac{M^2}{2\omega^2} + \text{terms in sine and cosine of } 2\omega t. \quad (25)$$

These results are now passed through a filter to remove the second harmonic terms. A filter having a transfer function $\gamma/(s+\gamma)$ will suffice. This filtering action is described by

$$\frac{1}{\gamma}\dot{u}_\alpha + u_\alpha = (a-\alpha)\frac{M^2}{2\omega^2} \quad (26)$$

$$\frac{1}{\gamma}\dot{u}_\beta + u_\beta = (b-\beta)\frac{M^2}{2\omega^2} \quad (27)$$

where u_α and u_β are the filter outputs. If we now set

$$u_\alpha = \frac{\dot{\alpha}}{K_\alpha}; \qquad u_\beta = \frac{\dot{\beta}}{K_\beta},$$

we obtain:

$$\ddot{\alpha} + \gamma\dot{\alpha} + \alpha\frac{K_\alpha M^2}{2\omega^2}\gamma = a\frac{K_\alpha M^2}{2\omega^2}\gamma \quad (28)$$

$$\ddot{\beta} + \gamma\dot{\beta} + \beta\frac{K_\alpha M^2}{2\omega^2}\gamma = b\frac{K_\alpha M^2}{2\omega^2}\gamma, \quad (29)$$

which are (12) and (13) in the body of the paper.

The final step is to integrate u_α and u_β obtaining

$$\alpha = K_\alpha\int u_\alpha dt \quad (30)$$

and

$$\beta = K_\beta\int u_\beta dt. \quad (31)$$

These signals are used to position servo-driven potentiometers which are used to synthesize B and $1/H'$.

Appendix II

List of Symbols

a_r = coefficients characteristic of the physical plant,
c_n, d_n = coefficients characteristic of the controller,
f = the force applied to the plant,
g = a general functional relationship,
B = the controller transfer function,
H = the plant transfer function,
H' = an approximation to H,
K_α, K_β = tracking loop gains,
M = the magnitude of a sinusoidal perturbation at the plant output,
P = the Laplace transform of the perturbing signal,
a, b = plant parameters, subject to change,
c, d = controller parameters, constants,
$e = x_0 - x$ = an error function,
s = the Laplacian variable,
u_α, u_β = filter outputs,
x = the plant output,
x_0 = the system input,
z = the tracking loop error function,
α, β = approximations to a and b,
γ = the filter parameter,
θ = the phase angle of the perturbation of the output,
ω = the frequency of the perturbation.

Bibliography

[1] R. F. Drenick and R. A. Shahbender/ "Adaptive servomechanisms," *Trans. AIEE*, vol. 76 (*Applications in Industry*), pp. 286–292; November, 1957.
[2] J. A. Aseltine, A. R. Mancini, and C. W. Sarture, "A survey of adaptive control systems," IRE Trans. on Automatic Control, vol. AC-6, pp. 102–109; December, 1958.
[3] M. Margolis and C. L. Leondes, "A parameter tracking for adaptive control systems," 1959 IRE WESCON Convention Record, pt. 4, pp. 104–116.
[4] M. Margolis, "On the Theory of Process Adaptive Control System, the Learning Model Approach," Ph.D. dissertation, University of California, Los Angeles; October, 1959.
[5] M. Margolis and C. L. Leondes, "On the Philosophy of Adaptive Control Systems," Dept. of Engrg., University of California, Los Angeles, Air Force Office of Scientific Res., TN-59-1199; January, 1960.

Adaptive Servo Tracking*

A. I. TALKIN†, MEMBER, IRE

Summary—**This paper describes a self-adaptive sampled-data radar tracking loop. The adaptive tracking loop may be considered to be a low-pass filter with a variable bandwidth. The loop is designed to adapt rapidly to changes in the input signal by monitoring both the apparent error and the loop output.**

Results show a mean tracking accuracy 25 to 34 per cent higher than that of a comparable linear system at a receiver SNR of 10 db. In other terms, to attain the same accuracy the comparable linear system requires a SNR of 13 db. This improvement in performance is obtained with relatively little additional circuitry. In general, the degree of improvement obtained by adaptive tracking will depend on the compromises made in the design of the unadapted loop to track the most difficult type of target.

I. Origin of Problem and Objective

THIS investigation developed as an offshoot of a track-while-scan radar project. The radar system was designed to obtain as high a tracking accuracy as possible, particularly in angle, in order to use the data for missile guidance. The angle and range tracking loop configurations are identical in this radar, with practically no mutual effects between them as long as the target is held in track by both. The work described was applied to angle tracking only, but the results are equally applicable to range tracking. The tracking loops include two operational amplifier integrators in a type II servo loop, yielding zero position error and zero velocity error in tracking a noise-free constant velocity signal. The design of the type II tracking loop involved specifying the roots of a characteristic equation of second order, or equivalently, specifying two arbitrary parameters. In the article by Smith, *et al.*,[1] which provided the basic tracking loop design, the parameters j and k were used rather than the roots themselves. The relation between the roots of the characteristic equation and the parameters j and k is given by

$$r_1, r_2 = \frac{4j^2 - 2k - 1 \pm \sqrt{(1 + 2k + 4j)(1 + 2k - 4j)}}{4j^2}.$$

It has been shown[2] that the tracking loop acceleration error coefficient equals $\tau^2 j^2$, where τ is the scan interval. The value of j is determined from the relation $j^2 = C/\tau^2\alpha$, where C is the tracking gatewidth[3] dictated by the radar beamwidth and the desired target resolution, and α is the effective target acceleration seen by the tracking loop for the most unfavorable target trajectory that is tactically feasible.[4] A particular target trajectory chosen for most of the experimental tests is described in Section IV. The value of k may be selected to give the best noise filtering or to minimize the error peak, when the target executes a turn. The article by Smith, *et al.*,[1] shows that for best noise filtering (for the case of white noise), $k = j$. However, this choice results in a higher peak error on the turns than the critically damped response (specified by $k = 2j - \frac{1}{2}$). The critically damped response was the one selected as a reasonable compromise for the design, since it is important to minimize servo error peaks on turns to prevent the combined noise and servo error from exceeding the tracking gatewidth, with a consequent high probability of breaking track entirely. If it were not for the restriction on the maximum acceleration lag $\tau^2 j^2 \alpha$, the noise response of the loop could be reduced by making j as large as possible while still maintaining the relation $k = 2j - \frac{1}{2}$. It is apparent that an adaptive system is possible by having j large until the target accelerates, then decreasing j as target acceleration increases, varying k simultaneously to maintain the relation $k = 2j - \frac{1}{2}$. This is equivalent to having a variable bandwidth filter that rests at a preset minimum bandwidth until signal accelerations force the bandwidth to increase. Such an adaptive filter offers the possibility of greater noise reduction and, consequently, higher tracking accuracy than the comparable system with fixed parameters. Experimental test results verify the above hypothesis. Data taken to compare the performance of the adapted and unadapted loops show that the adapted loop will track a target at 7-db SNR with the same integrated absolute error as that of the unadapted loop in tracking the same target at 10-db SNR. Details of how the performances of the two systems are compared are discussed in Section IV.

II. Previous Work in This Field

Although the design of adaptive systems is a relatively new field, there already exist a large number of papers[5] presenting complex analyses of the various aspects of adaptive systems. Adaptive loops that will follow polynomial signals of predetermined degree in the

* Received by the PGAC, January 30, 1961; revised manuscript received, March 10, 1961.

† Diamond Ordnance Fuze Labs., Washington, D. C.

[1] C. H. Smith, D. F. Lawden, and A. E. Bailey, "Characteristics of sampling servo systems" (*Proc. of the Cranfield Conf., 1951*), in "Automatic and Manual Control," A. Tustin, Ed., Academic Press, Inc., New York, N. Y.; 1951.

[2] A. I. Talkin, "Sampled-Data-Radar Tracking Analysis," Diamond Ordnance Fuze Lab., Washington, D. C., Rept. No. TR-661; March 16, 1959 (classified).

[3] The maximum error that can be tolerated before the probability of breaking track becomes high is given by the gatewidth.

[4] One such trajectory would occur for a target executing its maximum evasive maneuver, generally a 90° turn just before expected intercept by the defending missile.

[5] J. A. Aseltine, A. R. Mancini, and C. W. Sarture, "A survey of adaptive contro systems," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-6, pp. 102–108; December, 1958.

presence of Gaussian noise of known power spectrum are described by Drenick and Shahbender.[6] An acceleration estimator is designed based on the Zadeh-Ragazzini theory of prediction, and the control voltage output of the estimator is used to vary the main loop parameters to maintain an optimum relationship among them, which relationship is based on assuming steady-state conditions. In our problem, no steady state of constant acceleration can exist. It turns out that approximating the second derivative of the main loop output by cascading two RC filters is an adequate measure of signal acceleration. Certainly, from a practical standpoint this is simpler than designing and realizing an optimum predictor.

Roberts[7] has designed an adaptive, or self-optimizing control system based on the Wiener theory. That is, the power spectrum of the signal is assumed to be known. The optimum linear system is shown to be dependent on the ratio of the mean square levels of signal and noise (for white noise). When these levels are not known in advance or are slowly varying, the optimum adjustment may be maintained by forcing the apparent error[8] power spectrum to be white. This is accomplished by separating the low- and high-frequency components of the error and developing a control voltage whenever the absolute values of the low- and high-frequency error components are unequal. The treatment is largely concerned with self-optimization with constraints on the output acceleration. Such considerations are important where mechanical or electromechanical elements are involved. The tracking loop discussed in this report is an all-electronic one, and, hence, is unaffected by mechanical component limitations.

A basic principle of adaption, forcing the apparent error to be white when the noise is white[7] carries over to the problem treated herein. However, the technique by which it is accomplished is considerably different from that of Roberts,[7] basically because of the different input signal characteristics in the two problems.

III. Details of Present Approach

A. Technique for Operating Along a Locus

Fig. 1 represents the basic type II tracking loop that is the subject of this paper.

Fig. 2 is a reproduction from Smith, *et al.*,[1] showing typical step function responses for various choices of parameters j and k. Fig. 3 is a possible modification of the basic loop to permit operation with any effective values of j and k simply by varying the control voltages (M, N) into the three electronic multipliers.

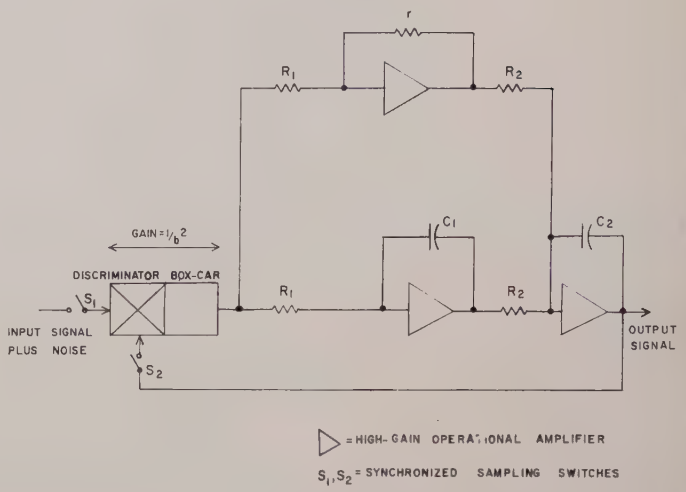


Fig. 1—The basic type II linear tracking loop.

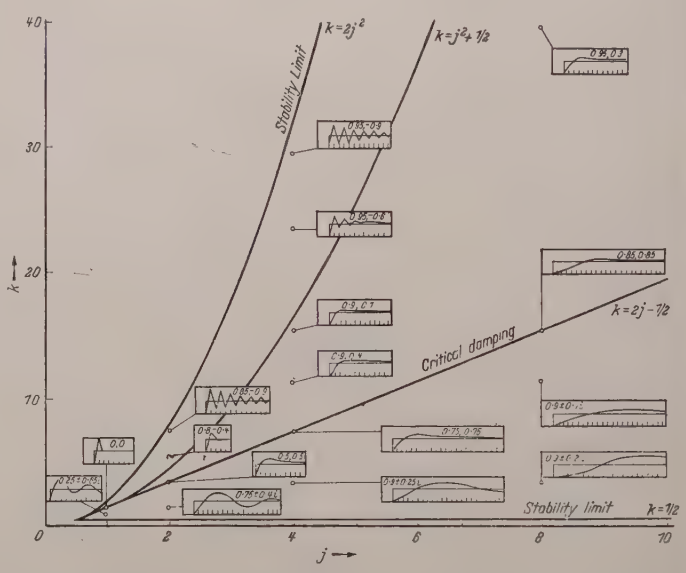


Fig. 2—Step response of two-stage sampling servo. Figures in inset diagrams are approximate values of roots of characteristic equation.

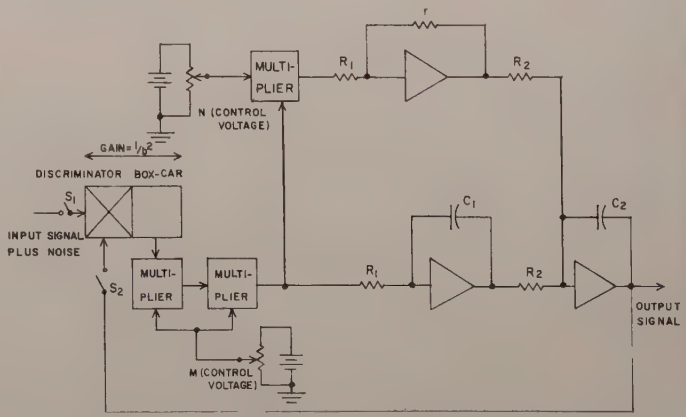


Fig. 3—Generalized loop for operation along any locus.

[6] R. F. Drenick and R. A. Shahbender, "Adaptive servomechanisms," *Trans. AIEE*, vol. 76 (*Applications in Industry*), pp. 286–292; November, 1957.

[7] A. P. Roberts, "Self-Optimising Control Systems for a Certain Class of Randomly Varying Inputs," Royal Aircraft Establishment, Farnborough, England, Tech. Note No. G.W. 507; January, 1959.

[8] The apparent error is the difference between the input signal plus noise and the loop output.

The relations between k and j and the circuit parameters in Fig. 1 are given by

$$j = \frac{b}{\tau}\sqrt{R_1C_1R_2C_2} \tag{1}$$

and

$$k = rC_{1/\tau} \tag{2}$$

where τ is the time between samples (500 msec in our system).

j and k may be varied by varying the effective values of b and r. Let j, k correspond to an initial operating point specified by (1) and (2). Define a capital J, K by the relations

$$J = j/M \quad \text{and} \quad K = Nk. \tag{3}$$

Then

$$J = \frac{b}{M\tau}\sqrt{R_1C_1R_2C_2}; \tag{4}$$

$$K = NrC_1/\tau. \tag{5}$$

To operate on the locus $K = 2J - \frac{1}{2}$,

$$NrC_1/\tau = \frac{2b}{M\tau}\sqrt{R_1C_1R_2C_2} - 1/2 \tag{6}$$

or

$$N = \frac{h}{M} - \frac{\tau}{2rC_1}, \tag{7}$$

where

$$h = \frac{2b\sqrt{R_1C_1R_2C_2}}{rC_1}. \tag{8}$$

Fig. 4 shows the loop configuration for operation along the specific locus, $K = 2J - \frac{1}{2}$. Note that only two multipliers are necessary. The multiplication by $1/M$ required by (7) is equivalent to taking part of the signal to the final integrator from the first multiplier.

B. Definition of Standard Target Trajectory

It was not considered realistic to represent the target trajectory of a single aircraft either by a power spectrum or a polynomial of known degree. Actually, the attacking aircraft will fly a relatively straight course, with a slight possibility of one evasive maneuver just before reaching the expected intercept range by a defensive missile. Thus the problem is to design an adaptive loop with a minimum bandwidth for tracking targets flying practically straight courses, and which has the capability of rapidly and continuously increasing its bandwidth to maintain track should the target go into an evasive maneuver or should acceleration become important because of the closing of the target range. As a result of these considerations, a standard

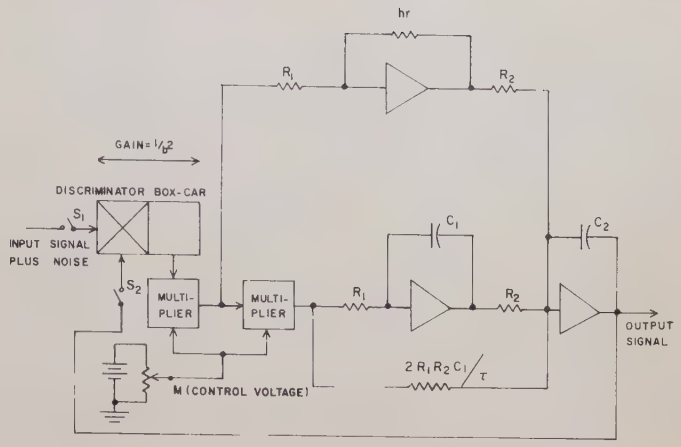


Fig. 4—Loop configuration for operation along locus $K = 2J - \frac{1}{2}$.

moving target trajectory was specified. The trajectory is that of an aircraft flying at 420 mph on an arc with a radius of 14 nautical miles from the radar. The target makes a 180° turn at each end of its trajectory, using its maximum turning acceleration of 4 g.

The target signal was generated by passing a 0.005-cps symmetrical triangular waveform through an RC filter of 0.02-cps corner frequency. The target then is moving at a constant angular rate of 0.44 degree per second for about 86 seconds, and then makes a U turn in another 15 seconds. The acceleration going into the turn is much greater than that coming out of the turn. This is considered to be a more realistic simulation of true air flight than a symmetrical turn.[9] It represents a greater challenge to the adaptive system since there is less warning of a turn available in the signal. Tracking error is measured by integrating the absolute value of the error over a period of 6 minutes and 44 seconds, or two complete round trips of the target. The input signal waveform is illustrated in Fig. 5.

C. Noise Spectrum and Scintillation

The noise that is superimposed on the echo pulses is receiver noise with about a 3.5-Mc bandwidth. The target signal for angle tracking is developed by boxcarring the return video burst consisting of $\frac{1}{2}$-μsec pulses with superimposed receiver noise. Since there can be but negligible correlation in the noise from pulse to pulse (since the pulses are 500 μsec apart), there cannot be any correlation in the noise from scan to scan (500 msec apart). Hence the correlation function of the noise is a delta function and the noise, as far as the tracking loop is concerned, can be considered white.

In addition to receiver noise, the tracking loop must cope with signal scintillation, a random variation in the amplitude of the return. If there were no receiver noise, such variations could be prevented from influencing the tracking by normalizing the error from the discrim-

[9] See W. M. James, N. B. Nichols, and R. S. Phillips, "Theory of Servomechanisms," M.I.T. Rad. Lab. Ser., McGraw-Hill Book Co., Inc., New York, N. Y., vol. 25, pp. 300–301; 1947.

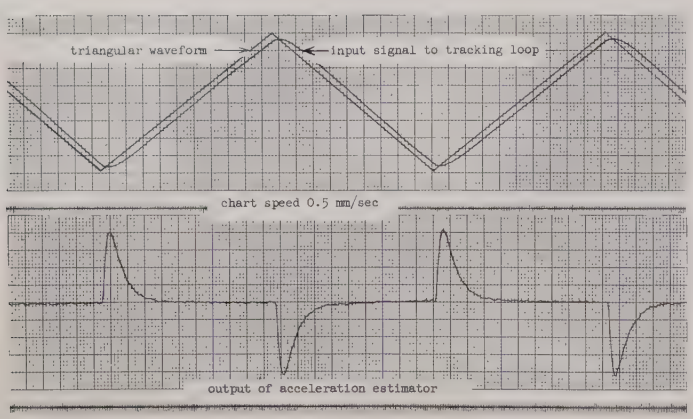


Fig. 5—Characteristics of input signal to the tracking loop.

inator. This has been done, but the normalization, of course, loses its effectiveness in the presence of receiver noise. Essentially, the effect of scintillation is to decrease the effective SNR. The data to be discussed in Section IV were taken for a nonscintillating target in the presence of receiver noise. Data taken with a scintillating target show the same absolute value reduction in integrated error by using an adapted loop, but the per cent reduction of error is less since the error for both adapted and unadapted tracking is increased.

D. Adaptive Technique (the Two Complementary Control Channels)

The adaptive schemes[6,7] examine either the servo input or the apparent error, processing one or the other of these signals through a filter (which may be fairly complex) or a computer, and using the filter or computer output to vary the parameters of the main loop. A more effective way to adapt in this problem proved to be by examining both the apparent error and main loop output.

The first approach was to examine only the apparent error with a smoothing filter and to make the main loop bandwidth proportional to the absolute value of the filter output. When the filter output was zero, the loop bandwidth was set at the minimum value required to maintain track in the presence of main loop integrator drifts and internal offsets. A limit was also set on the maximum bandwidth that the adapted loop may have. This limit was set somewhat below the theoretical upper limit of one half the sampling frequency. This was done to prevent single isolated noise pulses from causing unduly large errors. The adaptive system varies its j value from 2.5 to 8, corresponding to a bandwidth variation of about 3 to 1.

Experiments were made with a variety of impulse responses for the smoothing filter, but the improvement over unadapted tracking was marginal. This was primarily caused by the fact that the smoothing time of the filter had to be fairly short (2 or 3 seconds) so that the main loop bandwidth could increase rapidly enough to maintain track when the target went into its turn. As a result of the short smoothing time, the filter output was unduly noisy in the target straightaways.

The next experiment was to obtain an early indication of target acceleration by a double differentiation of the main loop output and to use this signal to increase main loop bandwidth. This does permit tracking to be maintained in the turn, but the acceleration signal decreases rapidly in the later or leveling-out portion of the turn. Hence, the bandwidth falls too rapidly with a consequent long-duration tail transient as illustrated in Fig. 6.

Only by using both the above techniques could significant reductions in tracking error be obtained. The bandwidth is now expanded when either or both of these complementary sensing or control channels produce an output. Since warning of acceleration was obtained from the main loop output, the smoothing filter time constant operating on the apparent servo error could be increased from 2 to 20 seconds. The complementing action of the two control channels is as follows. On entering the turn, the high-frequency loop increases the bandwidth to prevent losing track. After the initial sharp acceleration is over, the output from the high-frequency channel has gone to zero, but at the same time the low-frequency channel has had time to build up, thus rapidly eliminating the tail transient at the end of the turn. The reduced tail transient is illustrated in Fig. 7. Fig. 8 illustrates what happens when only the low-frequency control channel is active. Note the large initial error, but the rapid recovery.

The gains of the two control paths are determined experimentally by observing the servo error waveform in the turn and increasing the high-frequency loop gain to give a satisfactory initial error waveform, and increasing the low-frequency loop gain to give a satisfactory tail transient. These gain settings were not particularly critical,[10] but the lowest gain giving a satisfactory error waveform should be the best for performance in the presence of noise. Fig. 9 is a simplified schematic drawing showing the complete adaptive servo configuration.

IV. Results and Discussion

A. Basic for the Comparison of the Adapted with the Unadapted Loop

In comparing the performance of the adapted with that of the unadapted loop, the question arises how to select the parameter j of the unadapted system to have a fair comparison. It is both logical and convenient from a practical standpoint to select that j value for the unadapted loop that equalizes the integrated abso-

[10] A limited amount of data were taken showing that a 2-to-1 increase in gain of the low-frequency channel did not change the integrated absolute error at 10-db SNR.

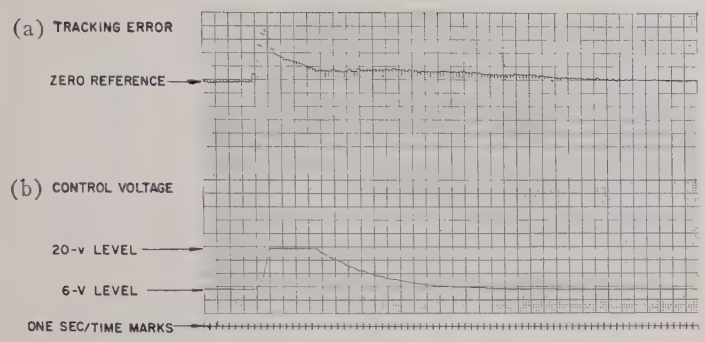


Fig. 6—Tracking the noise-free standard target through a turn, with only the high-frequency control channel active. [Note slow recovery in (a).]

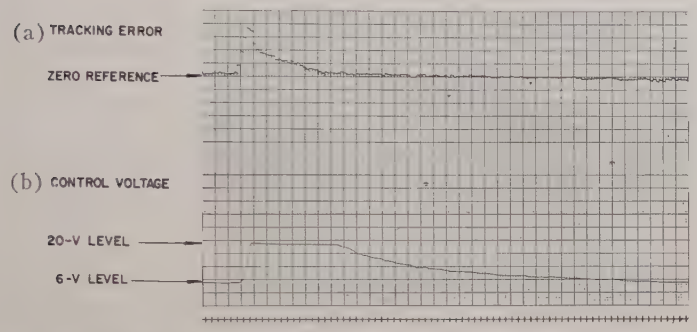


Fig. 7—Tracking through a turn, with both high- and low-frequency control channels active. [Note more rapid recovery in (a).]

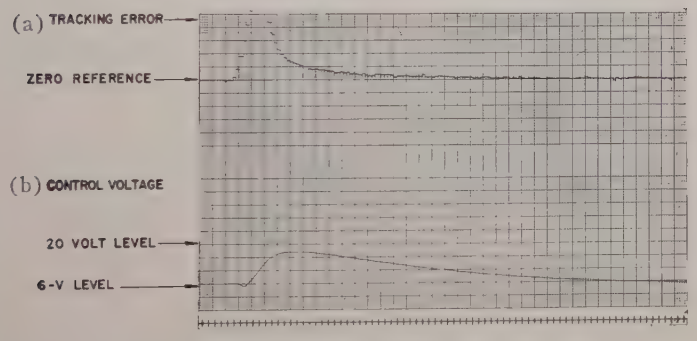


Fig. 8—Tracking through a turn with only the low-frequency control channel active. [Note large initial error in (a).]

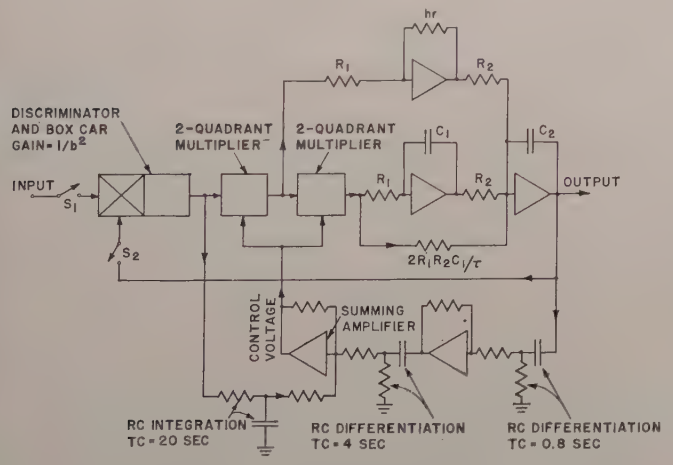


Fig. 9—Simplified schematic of complete adaptive servo.

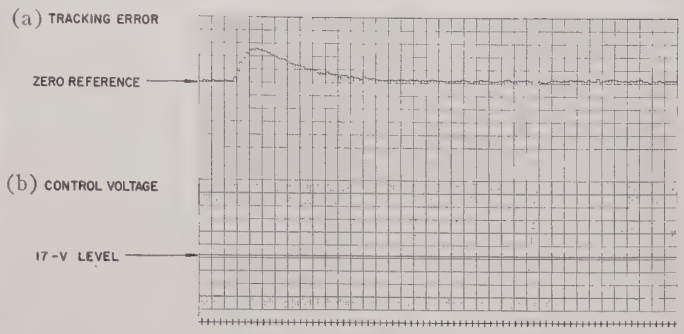


Fig. 10—Tracking through a turn, with unadapted loop. [Area of tracking error waveform adjusted via control voltage to equal that in Fig. 6(a).]

lute servo error of the two systems in tracking the noise-free standard moving target. This equalization was obtained at $j=3$. The error waveform in the turn for this unadapted loop is illustrated in Fig. 10.

The performances in the presence of noise may now be compared.

B. Results

1) *Standard Moving Target:* A 10-db SNR was chosen for most of the measurements with some additional data taken at 7-, 8-, 12-, and 13-db SNR. At 10-db SNR, the error of the adapted loop averages 25 per cent less than the unadapted loop error. Alternately, the adapted loop will track at 7-db SNR with the same average error as the unadapted loop at 10-db SNR, or the unadapted loop requires a 13-db SNR to track with the same error as the adapted loop at 10-db SNR.

2) *Stationary or Constant Velocity Target:* In tracking a stationary target at 10-db SNR, the adapted loop error is 34 per cent less than the unadapted loop error. This improvement should be maintained also in tracking a constant velocity target, since the basic loop is of type II, yielding zero servoing error for constant velocity signals.

An upper limit for the improvement that should be expected in this case is determined by the ratio of the bandwidth of the unadapted loop to the minimum bandwidth of the adapted loop. The formula is

$$\text{improvement (per cent)} = \left[1 - \sqrt{\frac{B_2}{B_1}}\right] \times 100,$$

where B_2 is the minimum bandwidth of the adapted loop, and B_1 is the bandwidth of the unadapted loop. This formula is easily derived, assuming white noise input and rectangular band-passes. The noise power output is then just proportional to bandwidth. Substituting in the formula the j ratios for the bandwidth ratios,

$$\frac{B_2}{B_1} \sim \frac{j_1}{j_2} = \frac{3}{8}$$

results in a maximum improvement of 40 per cent.

A summary of some typical data is given in Table I. The numbers given are the integrated absolute value of the errors over 6 minutes and 44 seconds, the time required for two complete target round trips. Only the relative values are significant; the absolute levels are determined by the gain of the error integrator.

TABLE I
RELATIVE TRACKING ERROR

A. Standard Moving Target

SNR	∞	13 db	10 db	7 db
Unadapted	30	70	92	—
Adapted	30	—	68	91
Manually adapted	—	—	69	—

B. Stationary Target

SNR	10 db
Unadapted	83
Adapted	55

3) *Manual Adaption:* In order to have some yardstick against which to measure the performance of the adaptive system in tracking the standard target, experiments were made with a manually adapted loop. That is, the loop bandwidth was switched to its maximum value ($j=2.5$) just before the target went into its turn, held at the maximum for the duration of the turn, and then switched back to its minimum value ($j=8$) for the straightaway. This experiment was carried out at 10-db SNR and gave no further reduction in error over the automatically adapted loop. Although a further reduction in error should be expected with the manually adapted loop, it is probably so slight that a considerable number of runs would be required to establish the amount of improvement.

4) *Hardware Required for Adaption:* Fig. 9 shows that two 2-quadrant multipliers and two operational amplifiers are required to adapt the main loop. The two multipliers are transistorized plug-ins with dimensions 4 inches by 2 inches by $1\frac{5}{8}$ inches. Their design has been presented by Ravilious.[11] Evidently, adding adaptivity to a radar tracking loop is not unduly costly.

5) *Applications:* The ideas used to produce an effective adaptive servo or variable bandwidth filter could be applied to any servo or filter that has to deal with a signal corrupted by noise, and where the upper frequency cutoff is set by the possibility that there may be high-frequency components in the signal that must be accepted. Radar tracking probably presents the greatest field for application of this type of adaptation. On radar systems where the tracking accuracy is adequate, the adaptive loop can yield the benefit of increased tracking range or increased immunity to wideband noise jamming.

CONCLUSION

It has been demonstrated experimentally that it is possible to increase the accuracy of a type II radar tracking loop by adaptive control of the bandwidth. The fast reaction required of the adaptive part of the loop can be obtained with no sign of instability in the main loop. In this investigation, main loop stability was secured by operating along the locus of critical damping.

[11] C. F. Ravilious, "Four-Quadrant Analog Voltage Multiplier," Diamond Ordnance Fuze Labs., Washington, D. C., Tech. Rept. No. 838; May 16, 1960.

Precision of Impulse-Response Identification Based on Short, Normal Operating Records*

R. B. KERR†, MEMBER IRE, AND W. H. SURBER, JR.†, MEMBER, IRE

Summary—The characteristics of a process may be estimated from an observation over a finite interval of time of the input and output variables of the system during a period of normal operation. The determination of the most effective method of analyzing the observed data and of estimating the probable errors in such an analysis is an important problem in the study of complex processes and in the design of adaptive controllers for time-varying systems. For adaptive control systems, it is desirable to base this system "identification" analysis on as short an operating record as possible, consistent with the specified degree of accuracy to be obtained. This paper is concerned with the problem of impulse-response estimation based on such short "normal operating" records.

If the measurements of the system variables are corrupted by noise, the impulse-response parameter estimates will be random variables, since, for a given record length, these estimates vary from one sample of the observed data to the next, depending on the variation in the characteristics of the noise and the input signal during each short segment of the record. The expected "integrated-squared-error" between the actual and the computed impulse responses is shown to depend only on the input signal and the noise characteristics. A method of computing the expected-integrated-squared error for a given input signal is developed to provide a test of the reliability of the identification routine for each analysis. With assumptions on the statistical nature of the input signal, this "sufficient signal" criterion is transformed to a "sufficient record length" criterion. Examples are given for two such specific assumptions:

1) An input signal with a Gaussian amplitude distribution.
2) A switching-type input signal which jumps between +1 and −1 with a random distribution of switching times. Results are presented in sampled-data form.

I. Introduction

THE problem of identifying the dynamic characteristics of a process by observing its input and output variables over a finite period of time during normal operation is a very important one in a number of fields. In the study of complex processes, where the determination of a suitable mathematical model for the system is the basic objective, relatively long record lengths may be used, thus enabling the signals to be described statistically. For an adaptive control system, where the controller is operating on a real time scale, it is very desirable to base the identification program on the observation of the system operation over as short a time interval as possible, consistent with the desired degree of accuracy in the estimation of the system parameters. This is particularly true if the system characteristics are varying with time and a time-invariant model is to be used to represent it approximately.

The primary objective of the process-identification program is the estimation of a set of parameters which will enable the future behavior of the system in response to its input or control signal to be predicted satisfactorily. If this is the case, then sufficient information is available to design the controller, *i.e.*, enable the proper value of the manipulated variable to be determined as a function of the input signal and the current state of the process. If the characteristics of the system are varying sufficiently slowly with time, relative to its "memory" time or settling time T_s, and if sufficiently "short" observation and prediction intervals are used, the system model may be assumed to be time-invariant. The system is then represented by a sequence of such models, the parameters being recomputed at intervals. It is also desirable to base the parameter estimates on records of the "normal" operating signals. There may be periods, however, when the nature of the normal control signals, due to either low amplitudes or certain waveform properties, are such that a sufficiently accurate estimate of the parameters cannot be made from the desired observation interval. It should be possible to detect these cases by a suitable computer routine in order to prevent the use of unreliable data by the controller. The parameter-estimation routine could then either by automatically extended over a longer interval, or special test signals could be deliberately injected. The discussion in this paper will be restricted to the use of normal operating records for parameter estimation.

For short-duration records, a strictly statistical description of the input and output signals is precluded. A conflict of requirements arises, in fact, since it is desirable to use as long a record as possible for noise smoothing, but as short a record as possible so that the system may be assumed to be time-invariant over the estimation interval. Hence, for a given rate of parameter variation, a given noise level, and a given type of control signal variation, there should exist an optimum record length if an approximate time-invariant model is to be used to represent the system for a definite time into the future, *i.e.*, until the results of the next computation of parameters are available.

* Received by the PGAC, December 15, 1960; revised manuscript received, March 13, 1961. This paper is based upon research work sponsored by the Aeronautical Research Associates of Princeton, Inc., Princeton, N. J., for General Precision Inc., New York, N. Y.

† Dept. of Elec. Engrg., Princeton University, Princeton, N. J.

II. Assumptions and Sources of Error

The type of system to be analyzed is indicated in Fig. 1. For a control system, where the magnitude of the input forcing function is normally limited by some form of saturation, it is assumed that the effective control signal after the limiter is available for observation. A suitable transfer function for the linear part of the system is to be determined. The effects of random disturbances applied throughout the process and of measurement errors in observing the output are represented by the equivalent noise generator $n(t)$, which is mixed with the ideal output signal.

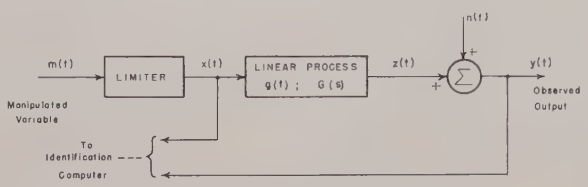


Fig. 1—System to be analyzed.

The system equations for this configuration are

$$z(t) = \int_{-\infty}^{t} x(\tau) \cdot g(t - \tau) d\tau, \tag{1}$$

and

$$y(t) = z(t) + n(t), \tag{2}$$

where

$g(t)$ = impulse response of the linear part of the system,
$x(t)$ = input signal to linear part of the system,
$z(t)$ = output variable in the absence of noise,
$n(t)$ = equivalent noise signal,
$y(t)$ = observed output variable.

There are a number of different types of mathematical models that could be selected to represent the dynamic process characteristics. The identification routine then estimates the values of the parameters for the given model to obtain the "best" fit to the observed noise-corrupted data in some sense, such as a least-squares fit between the observed system trajectory in its phase space and the model trajectory for a given input signal. Certain types of models, however, impose very stringent requirements on the behavior of the output variable from the model. If the model were assumed to be a second-order linear differential equation, for example, even the best choice of the model parameters might result in an unsatisfactory representation of a higher-order system. The impulse response characteristic $g(t)$, by contrast, seems to be a very suitable form of linear system representation for the study of the effects of noise on the parameter estimates. The advantages of the $g(t)$ model are, first, that it involves the minimum number of assumptions concerning a specific structure for the model and, second, the relative ease with which other equivalent or simpler sets of parameters characterizing the system dynamic response can be computed from $g(t)$. The number of parameters required and their accuracy depend, of course, on the type of controller that is to be used and the degree to which it is desired to approach optimum performance.

In order to describe $g(t)$ by a finite number of parameters, it will be assumed that the system has a finite settling time T_s, beyond which the $g(t)$ function is negligibly small, and that the system has a finite significant bandwidth.

The sources of error in the impulse-response identification may be classified as follows.

A. Noise and Measurement Errors

The effects of noise, as represented by $n(t)$, on the accuracy of parameter estimation will be discussed in detail in the remainder of this paper. As the length of the observed record is increased, thus increasing the number of effectively independent data samples, the precision with which the set of parameters is estimated increases. This type of noise smoothing is a function of the degree of redundancy in the data, *i.e.*, the number of independent output data samples relative to the number of parameters to be estimated.

B. Model Structure Errors

1) *Finite Settling-Time Assumption:* If the assumed settling time T_s is too long, the apparent number of parameters to be estimated will be increased. In addition to the increased computer capacity required, the decrease in redundancy for a given observation interval decreases the effective noise smoothing and results in poorer precision. If T_s is too short, however, a systematic error will be introduced into the estimates of the retained parameters. The best choice of T_s for a fixed measurement interval T_m would be obtained by reducing T_s, thus increasing the apparent redundancy of the observed data and decreasing the fluctuations caused by noise, until the systematic error became of the same order of magnitude as the expected noise-induced estimation error.

2) *Finite Bandwidth Assumption:* If the system is assumed to have an effective upper cutoff frequency f_c above which the output $z(t)$ will not respond significantly (*i.e.*, the energy in the components of z above this frequency is negligible relative to the noise energy), then $g(t)$ may be represented by a set of values of $g(t)$ separated at intervals of time $\Delta t = [1/(2f_c)]$ with a negligible loss of information. If Δt is chosen to be smaller than this, due to our initial ignorance of the actual system characteristics, then a larger number of parameters will apparently be required for a given T_s, since $N = (T_s/\Delta t)$ is the number of g_k values which must be estimated. In addition to requiring a faster sampling rate and a greater computer capacity, this increases the expected error in the parameter estimates, for white noise, essentially in proportion to the square root of the increase in N, for a given measurement interval T_m. Although the redundancy in the observed data remains almost constant, being determined primarily by the ratio of T_m to T_s (which does not change as Δt is varied), the change of the output variable $z(t)$ becomes relatively less from sample to sample, thus increasing the effect of the noise fluctuations. A quantitative example is given in the Appendix for a very simple structure of the input signal $x(t)$. The same conclusion can be reached using another approach, by observing that the short Δt interval allows less filtering of the observed output $y(t)$ in determining the values of the output data samples, thus increasing the effect of the noise energy relative to the signal energy. This effect is very closely analogous to the problem of the choice of a Δf interval in the estimation of the frequency response characteristics of the system, a problem which has been extensively discussed in the literature.[1]

If Δt were to be increased beyond the $[1/(2f_c)]$ interval, this increase would introduce a systematic structural error in the determination of the equivalent model, since the fine structure of the system $g(t)$, *i.e.*, the high-frequency response characteristics, would be smoothed out and, therefore, lost. The expected error in the parameter estimates due to noise induced fluctuations would be decreased, however. Again, as in (1), a "best" choice of Δt might be considered to be one which would approximately equalize the noise-induced and the systematic expected errors.

3) *Time Variation of the System Parameters:* It is assumed in the analysis in later sections that the process parameters are constant over the measurement interval T_m. The controller, for an adaptive system using this identification routine, would also assume that these estimated parameters remain time-invariant over the prediction interval as well. The actual parameters must change sufficiently slowly with time relative to the degree of accuracy required by the controller for this assumption to be valid.

C. Sampling Time and Computation Time Delays

If the system variables are coupled to the identification computer through analog-to-digital conversion transducers and sampled at intervals of Δt, an effective sampling delay of $(\Delta t/2)$ is introduced by the data acquisition system. An additional computation time interval would elapse before the controller received each set of results.

The effective sampling delay $(\Delta t/2)$ is due to the need to filter, or smooth, the actual signals before sampling in order to avoid frequency aliasing with a consequent increase in the noise fluctuations per sample. Although this delay could be avoided by postulating an idealized sampled-data system, using an impulse train for data sampling, this might be highly undesirable in practice.

One method of obtaining the samples of $y(t)$ would be to average y over each interval Δt as indicated in

$$y_k = \frac{1}{\Delta t}\int_{(t_k-\Delta t)}^{t_k} y(t)dt, \tag{3}$$

so that y_k represents an estimate of $y(t)$ delayed by $(\frac{1}{2}\Delta t)$ behind the time t_k at which the sample becomes available to the computer.

For a finite settling time T_s and a finite sampling time Δt, (1) may be modified as follows:

$$z(t) = \int_{(t-T_s)}^{t} x(\tau)\cdot g(t-\tau)d\tau = \int_0^{T_s} x(t-\tau)g(\tau)d\tau, \tag{1a}$$

or, in sampled-data form,

$$z_k = \sum_{n=0}^{N-1} x_{(k-n)}\cdot g_n, \tag{4}$$

where

$$g_n = \int_{n\Delta t}^{(n+1)\Delta t} g(t)dt \tag{5}$$

and

$$y_k = z_k + n_k. \tag{6}$$

For the sampled-data formulation of the identification problem, for which the number of impulse-response parameters to be estimated is $N = (T_s/\Delta t)$, the minimum number of input samples observed must clearly be $[N+(N-1)] = (2N-1)$, in order that a complete set of N equations be formed from (4), as indicated in Fig. 2.

[1] R. B. Blackman and J. W. Tukey, "The measurement of power spectra from the point of view of communications engineering," *Bell Sys. Tech. J.*, vol. 37, pp. 185–282, January; pp. 485–569, March, 1958.

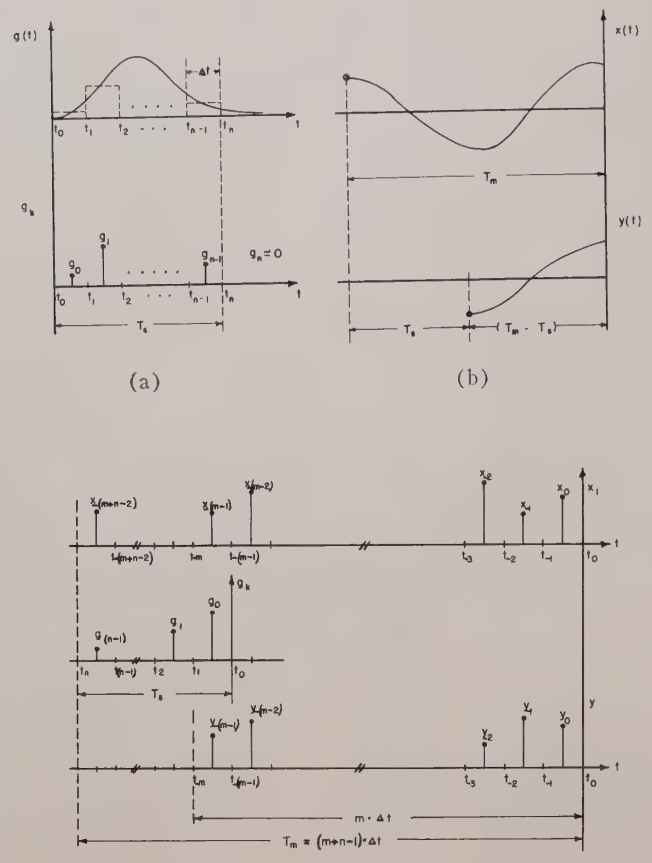


Fig. 2—Sampled-data notation. (a) Impulse response. (b) Continuous signal records. (c) Sampled-data sequences: x_k and y_k are available to computer at time t_k.

Any additional data samples observed, by using a longer T_m interval, will allow the formation of an excess number of equations relative to the number of independent parameters to be computed. This redundancy can be used for noise smoothing, as outlined in the following section. It is convenient to define a "redundancy factor," R, as follows:

$$R = \frac{r}{N} = \left(\frac{M - N}{N}\right) = \left[\left(\frac{M}{N}\right) - 1\right], \qquad (7)$$

where

$N = (T_s/\Delta t)$ = minimum number of output samples,
$(2N-1)$ = minimum number of input samples,
M = total number of observed output samples,
$r = (M-N)$ = number of excess output samples.

If M and N are expressed in terms of time intervals, then the equation for R becomes,

$$R = \left[\left(\frac{T_m - 2T_s}{T_s}\right) + \frac{1}{N}\right]. \qquad (7a)$$

For continuous signals with no bandwidth limitation and therefore no time quantization due to sampling, Δ is zero, N is infinite, and the minimum observation interval for zero redundancy becomes $T_m = 2T_s$. For a bandwidth-limited system, however, it is interesting to note from (7a) that the minimum observation time to obtain a nonredundant set of data points is less than $2T_s$, and has the form

$$(T_m)_{\text{minimum}} = \left[\left(2 - \frac{1}{N}\right) \cdot T_s\right],$$

so that it would be equal to T_s if N were one. If the linear system were an amplifier with zero memory time, N would be one, since the single parameter would be the amplifier gain. In this case $T_m = \Delta t$, the minimum sampling time required to observe one data point, and since T_s also equals Δt using the relation $T_s = (N \cdot \Delta t) = \Delta t$, the minimum $T_m = T_s$. For a zero memory system, of course, T_s would be zero ideally, as would also be the case for T_m. The minimum bound of Δt is set for both T_s and T_m by the assumed mechanism of forming each data sample by averaging over a single Δt interval.

III. Least-Squares Estimation of Impulse Response

Levin[2] has pointed out that a least-squares estimate of impulse response, based upon short operating records and the model of Fig. 1, follows essentially the same procedure as the statistical filter design procedure of Wiener[3] and others. That is, one calculates as an estimate of $g(\tau)$ that $\hat{g}(\tau)$ which minimizes E^2, where

$$E^2 \triangleq \int_a^b [z(t) - y(t)]^2 dt$$

$$= \int_a^b \left[\int_0^{T_s} g(\tau)x(t-\tau)d\tau - y(t) \right]^2 dt. \tag{8}$$

Here, T_s is again the "settling time" of the system, and $y(t)$ is assumed to be observed over the interval a to b. It will be assumed at present that a good estimate for T_s may be made *a priori*. By applying standard variational techniques,[4] $\hat{g}(\theta)$ is found to satisfy

$$\int_0^{T_s} \hat{g}(\theta)\phi_{xx}(\tau, \theta)d\theta = \phi_{xy}(\tau), \tag{9}$$

where

$$\phi_{xx}(\tau, \theta) \triangleq \int_a^b x(t-\tau)x(t-\theta)dt \tag{10}$$

$$\phi_{xy}(\tau) \triangleq \int_a^b y(t)x(t-\tau)dt \tag{11}$$

$$0 \leq \tau \leq T_s; \qquad 0 \leq \theta \leq T_s; \qquad a \leq t \leq b.$$

The functions ϕ_{xx} and ϕ_{xy}, which might be termed "quasi-correlation" functions, bear a close resemblance to the usual statistical correlation functions. That is, if x is a continuous random function of time, its autocorrelation function may be defined as

$$\psi_{xx}(\tau - \theta) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{+T} x(t-\tau)x(t-\theta)dt.$$

ψ_{xx} gives the statistical properties of the random variable x and may be used to predict its future behavior. In our case, however, x is not a random variable and in fact is assumed to be completely unknown outside of the limited interval of observation. Hence ϕ_{xx} and ϕ_{xy} are not correlation functions at all, in the true statistical sense, and we have formulated the problem as a *regression* problem rather than as one to which correlation theory is applicable.[5]

In sampled-data formulation, (9) becomes (with capital letters denoting matrices, small letters vectors, and* denoting the transpose of a matrix)

$$\Phi_{xx}\hat{g} = \phi_{xy}, \tag{12}$$

or

$$\hat{g} = \Phi_{xx}^{-1}\phi_{xy}, \tag{13}$$

where

$$\Phi_{xx} = X^*X \tag{14}$$

$$\phi_{xy} = X^*y, \tag{15}$$

and

$$X = \begin{bmatrix} x_0 & x_{-1} & \cdots & x_{-(N-1)} \\ x_{-1} & x_{-2} & \cdots & \cdot \\ x_{-2} & x_{-3} & \cdots & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ x_{-(M-1)} & \cdots & & \cdot \end{bmatrix}, \tag{16}$$

M being the number of samples in the interval a to b, and N the number of samples in an interval of duration T_s. Hence,

$$\hat{g} = (X^*X)^{-1}X^*y, \tag{17}$$

which reduces, for a "minimum" record of $x(t)$, to

$$\hat{g} = X^{-1}y, \tag{18}$$

the solution for the "nonredundant data" case.

This procedure then gives a method for making use of redundant data in a systematic way. If the noise $n(t)$ is white and Gaussian, the method is optimum in the sense that it gives a maximum-likelihood estimate of g, which is efficient. If the noise is Gaussian but not white, the least-squares technique does not yield an efficient estimate. The maximum-likelihood estimate in this case is the Markov estimate, which is efficient[6] but which requires much more computing capacity and makes use of the spectral information concerning the noise.

It should be noted that although the estimation computer is required to multiply two $M \times N$ matrices, it need invert only an $N \times N$ matrix, where N is the number of samples of the estimated impulse response.

[2] M. J. Levin, "Optimum estimation of impulse response in the presence of noise," IRE Trans. on Circuit Theory, vol. CT-7, pp. 50–56; March, 1960.

[3] N. Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series," John Wiley and Sons, Inc., New York, N. Y.; 1949.

[4] R. Courant, "Differential and Integral Calculus," Nordemann Publishing Co., New York, N. Y., vol. 2; 1936.

[5] A. M. Mood, "Introduction to the Theory of Statistics," McGraw-Hill Book Co., Inc., New York, N. Y.; 1950.

[6] U. Grenander and M. Rosenblatt, "Statistical Analysis of Stationary Time Series," John Wiley and Sons, Inc., New York, N. Y., pp. 86–90; 1957.

IV. The "Sufficient Test Signal" Problem

The reliability of the estimate (17) will be seen to depend critically upon the nature (amplitude and waveform) of the particular input $x(t)$ upon which the estimate is based. A logical choice for a measure of this reliability would seem to be the expectation of the integrated-squared-error between the actual and the estimated impulse responses. An expression for the expected integrated-squared-error in the sampled-data formulation may be obtained as follows.

The actual output vector is

$$y = Xg + n, \tag{19}$$

where n is the noise vector made up of samples of $n(t)$. From (12),

$$\phi_{xy} = X^*y = \Phi_{xx}\hat{g}.$$

Hence,

$$X^*Xg + X^*n = \Phi_{xx}\hat{g}. \tag{20}$$

Since

$$X^*X = \Phi_{xx},$$

we have

$$\Phi_{xx}(\hat{g} - g) = X^*n, \tag{21}$$

or

$$(\hat{g} - g) = \Phi_{xx}^{-1}X^*n. \tag{22}$$

The total squared error will then be the sum of the squares of the elements of $\hat{g}-g$, or

$$I^2 = (\hat{g} - g)^*(\hat{g} - g). \tag{23}$$

If the noise is white (introduces independent random fluctuations at each sample point) and has a variance σ_n^2, then

$$\langle I^2 \rangle = \sigma_n^2 \sum_{i,j} a_{ij}^2, \tag{24}$$

where a_{ij} are the elements of the matrix

$$\Phi_{xx}^{-1}X^* = (X^*X)^{-1}X^*.$$

For a minimum record, this reduces to the matrix X^{-1}. We may normalize $\langle I^2 \rangle$ with respect to the signal variance σ_s^2 by defining $\alpha_{ij}=\sigma_s a_{ij}$. Then

$$\langle I^2 \rangle = \left(\frac{\sigma_n}{\sigma_s}\right)^2 \sum_{i,j} \alpha_{ij}^2, \tag{24a}$$

where

$$\sum_{i,j} \alpha_{ij}^2$$

may be interpreted as a "signal structure" factor.

Hence, we might define a "*sufficient*" input signal as one for which some specified $\langle I^2 \rangle$ is not exceeded, as this expected error is seen to depend *only* upon the input signal and the noise variance. Whether or not this is the best possible sufficiency criterion remains a question. It would seem that to answer this question, one would have to consider, for example, the over-all control system of which the system estimator is a part, and choose a criterion based upon some optimum over-all behavior of the system. However, if one does not know exactly how the controller is going to "use" the impulse-response information (and different controllers might use it in different ways), then the expected integrated-squared-error criterion would seem to be a logical choice.

The expected integrated-squared-error is seen to depend critically upon the determinant of Φ_{xx} (or X for the nonredundant record case). In particular, this determinant may vanish, yielding an infinite expected error. For example, if a nonredundant record is used, any linearly dependent rows or columns in the determinant of X, which indicate "periodicities" (in a sense) in x, yield a vanishing determinant. With redundant data, however, some dependence may be tolerated. For example, if

$$X = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 6 \end{bmatrix},$$

the first and third rows are linearly dependent, and yet the determinant of $\Phi_{xx}=X^*X$ does not vanish.

Hence, the nonredundant record is far more critical, so far as the determinant vanishing is concerned, than the longer redundant record. In fact, as the amount of redundancy increases, the determinant of Φ_{xx} becomes less and less critically dependent upon the particular x vector, as will be evident in Section V. It is noted that the determinant of Φ_{xx} may be small, due to either the "wave-shape" represented by x, or the elements of x themselves being small, which corresponds to the $x(t)$ signal having insufficient amplitude.

It is apparent that in a practical application, such as an adaptive control system, an expected-error computer could be included with the system estimation computer, as indicated in Fig. 3. The control computer would then receive information regarding not only the model parameters, but also the expected errors in the estimates of these parameters. The control routine might then base the computation of the control variable on both of these sets of information.

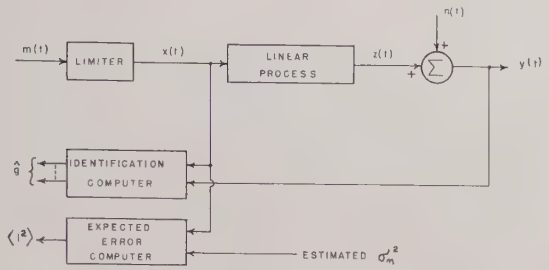


Fig. 3—Parameter confidence computation.

V. Dependence upon Record Length

A. A "Sufficient Record Length" Criterion

If the structure of the input signal itself or its statistics are known *a priori*, then it is possible to convert our "sufficient test signal" criterion into a "sufficient record length" criterion. For example, if the statistics of the input signal are known, we may determine the mean and variance of the expected integrated-squared-error as a function of the record length. That is, we may consider any particular input signal to be a member of the ensemble of signals of that particular record length, and average the expected integrated-squared-error over this ensemble. Hence, we are really doing two averagings of the error—one with respect to the ensemble of noise signals, and one with respect to the ensemble of input signals. This would therefore allow us to say with a certain degree of assurance, that with a record of a certain length we would have obtained a sufficient test signal. This in turn might permit the omission of the error computer in Fig. 3.

B. Examples of Record Length Dependence

An example is presented (Table I, next page) where x is made up of independent samples of a random variable $x(t)$ having a Gaussian amplitude distribution with variance one and mean zero.

The expected integrated-squared-error *per unit noise variance* was calculated for 20 sample records, each of several different record lengths. The settling time T_s of the system was assumed to extend over 6 of the sampling intervals, which corresponds to 6 samples of the impulse response being obtained. Table I shows the expected error values obtained for redundancy factors of 0, $\frac{1}{2}$, 1, 2, 4 and 8.

In Figs. 4 and 5, the mean and variance of the expected error are shown as a function of the redundancy factor. While the possibility of a "wild" result exists for a redundancy factor of 8, the probability of such an occurrence is slight—whereas for the nonredundant case, one would definitely need to calculate the expected error for each individual x vector in order to determine whether or not that particular vector was a sufficient test signal.

Hence it is clear that whether or not the error computer may be dispensed with depends very critically upon the available record length (which in turn depends upon the rate at which the system is time varying and the speed of data-reduction desired) and the allowable uncertainty in the estimate.

In Table II and Figs. 4 and 5 are shown the corresponding results for an x signal made up of independent samples of amplitude ± 1, each occurring with probability $\frac{1}{2}$ in each sampling interval. This corresponds to samples of a "bang-bang" type of input signal, having random switching times, as indicated in Fig. 6. The various probabilities for the number of switches in the interval T_s are given by the binomial distributions indicated in Fig. 6.

It should be noted that the two signals (Gaussian amplitude and random switching) have identical autocorrelation functions and, hence, the same power density spectra (Fig. 6).

VI. Conclusions

Short records of the normal operating signals may be used to estimate in a least-squares sense the impulse response of a linear system, to within an expected integrated-squared-error which is a function only of the noise and the input signal. If such an identification computer were incorporated in an adaptive control system, it would be desirable to also compute the error to determine whether or not the particular signal segment upon which the estimation was based was a "sufficient" test signal. A method of doing this has been developed. If the statistics of the input signal are known, this error computer may be eliminated for sufficiently long records, depending upon the confidence required in the estimate. For typical input signals, it is apparent that as the record length is increased, the expected integrated-squared-error decreases very rapidly—*i.e.*, a large increase in precision is gained from a small amount of redundant data. For a time-varying system, this increase in precision must be balanced against the increase in error due to assuming the system to be time-invariant over the longer record, in a way which has yet to be studied in detail.

For the same signal and noise variances (giving the same expected signal and noise energies over each measurement interval), the $\pm$ switching signal provides a

TABLE I

$(\sigma_s/\sigma_n)^2\langle I^2\rangle$ vs Record Length—Gaussian Amplitude Signal, $N = 6$

Redundancy R =	0	1/2	1	2	4	8
Mean $(\sigma_s/\sigma_n)^2\langle I^2\rangle$ =	40.42	1.98	1.02	0.601	0.305	0.142
Variance of $(\sigma_s/\sigma_n)^2\langle I^2\rangle$ =	14,217.8	1.44	0.26	0.0413	0.008	0.00045
Typical runs for $(\sigma_s/\sigma_n)^2\langle I^2\rangle$						
1	4.935	0.958	0.595	0.625	0.212	0.147
2	547.917	1.195	1.249	0.917	0.266	0.132
3	1.710	1.716	0.547	0.441	0.333	0.137
4	3.144	2.735	1.051	0.498	0.244	0.161
5	7.865	1.185	1.318	0.567	0.538	0.122
6	1.687	2.693	0.961	0.491	0.275	0.127
7	11.883	1.016	0.638	0.398	0.315	0.177
8	3.636	2.568	0.731	1.063	0.249	0.123
9	41.112	1.580	0.672	0.508	0.276	0.147
10	7.280	2.388	0.772	0.457	0.214	0.187
11	3.923	1.254	0.720	0.809	0.201	0.126
12	6.855	1.953	1.234	0.724	0.267	0.164
13	6.614	0.970	0.550	0.410	0.321	0.122
14	3.218	1.224	0.933	0.655	0.270	0.142
15	119.832	1.891	2.293	0.374	0.247	0.173
16	10.570	1.141	0.712	0.468	0.303	0.132
17	12.951	1.390	1.159	0.934	0.473	0.131
18	2.190	6.397	0.191	0.472	0.470	0.161
19	5.991	2.583	0.912	0.829	0.252	0.117
20	5.132	2.798	2.521	0.384	0.362	0.113

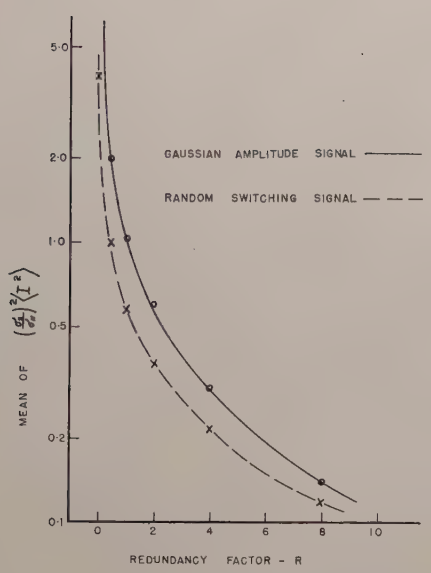


Fig. 4—Expected error mean vs redundancy.

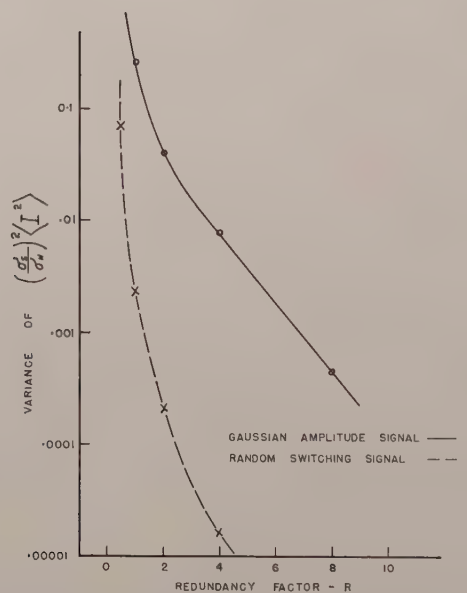


Fig. 5—Expected error variance vs redundancy.

TABLE II

$(\sigma_s/\sigma_n)^2\langle I^2\rangle$ vs Record Length—Random Switching Signal, $N = 6$

Redundancy R =	0	1/2	1	2	4	8
Mean $(\sigma_s/\sigma_n)^2\langle I^2\rangle$ =	3.928	1.002	0.576	0.375	0.215	0.118
Variance of $(\sigma_s/\sigma_n)^2\langle I^2\rangle$ =	3.451	0.0729	0.00246	0.000216	0.0000171	0.00000117
Typical runs for $(\sigma_s/\sigma_n)^2\langle I^2\rangle$						
1	5.000	0.828	0.586	0.379	0.212	0.120
2	3.750	1.511	0.542	0.378	0.213	0.117
3	3.250	0.952	0.544	0.356	0.215	0.118
4	3.750	0.792	0.544	0.402	0.214	0.118
5	6.000	0.871	0.542	0.370	0.212	0.120
6	3.750	0.828	0.546	0.380	0.214	0.119
7	1.500	1.147	0.542	0.358	0.212	0.117
8	2.500	1.511	0.612	0.364	0.217	0.117
9	6.000	0.792	0.542	0.387	0.210	0.120
10	1.500	0.780	0.673	0.356	0.223	0.119
11	4.500	0.828	0.542	0.420	0.211	0.118
12	2.500	0.875	0.647	0.370	0.217	0.118
13	1.500	1.511	0.542	0.387	0.210	0.118
14	5.500	0.896	0.674	0.370	0.223	0.120
15	4.500	0.792	0.542	0.356	0.211	0.119
16	3.750	1.011	0.612	0.380	0.217	0.118
17	3.250	0.828	0.542	0.364	0.213	0.117
18	3.297	1.511	0.561	0.374	0.216	0.119
19	6.000	0.952	0.542	0.363	0.219	0.120
20	3.750	0.792	0.542	0.383	0.210	0.118

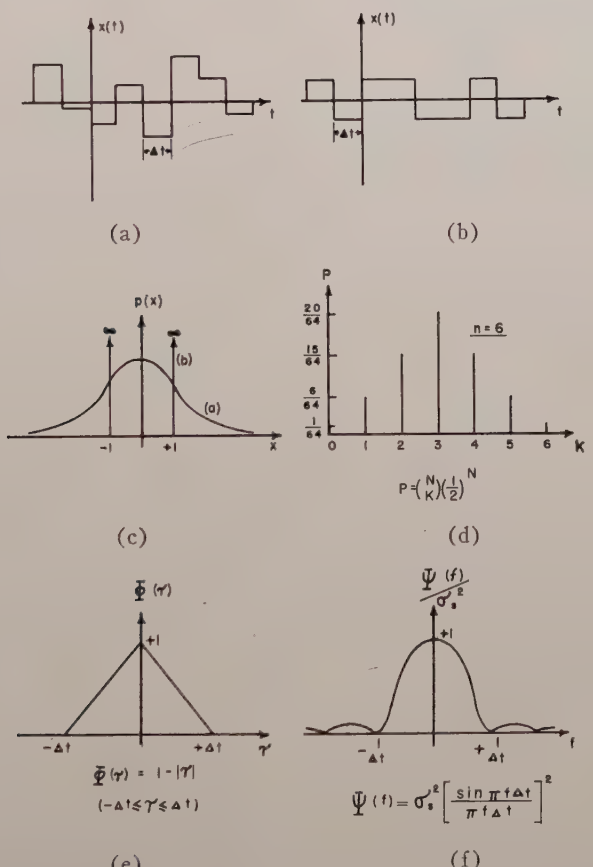


Fig. 6—Input signal characteristics. (a) Gaussian amplitude signal. (b) Random switching signal. (c) Amplitude densities of (a) and (b). (d) Probability P of K switches in the interval T_s. (e) Autocorrelation function of (a) and (b). (f) Power density spectrum of (a) and (b).

considerably smaller mean value for the expected integrated-squared-error $\langle I^2 \rangle$ in the estimated impulse response than a signal having a Gaussian amplitude distribution. It also results in a much smaller variance for $\langle I^2 \rangle$, and this variance falls off with increasing redundancy far more rapidly than for the Gaussian signal.

Appendix

Variation of System Bandwidth

The effect of increasing the assumed system bandwidth for a fixed measurement interval is to increase the expected integrated-squared-error in the impulse response, even though the degree of redundancy remains almost constant. This can be easily shown by using a very simple assumed structure for the input signal $x(t)$ as follows:

Let $x(t)$ be a periodic square wave, jumping between $+\sigma_s$ and $-\sigma_s$ at T_s intervals, and thus having a period of $2T_s$. Then it is easily shown from (24a) that for the nonredundant data case, the expected integrated-squared-error varies as

$$\left[\frac{\langle I^2 \rangle}{\sigma_n^2/\sigma_s^2}\right] \begin{cases} = 1 & \text{for } N = 1 \\ = (N/2) & \text{for } N \geq 2 \end{cases}, \tag{25}$$

so that $\langle I^2 \rangle$ increases directly with N, the number of parameters required to specify $g(t)$ over the T_s interval. For the redundant data case, and for *integer values of* R, it can also be shown that

$$\left[\frac{\langle I^2 \rangle}{\sigma_n^2/\sigma_s^2}\right] \begin{cases} = \left(\dfrac{1}{R+1}\right) & \text{for } N = 1 \\ = \dfrac{N}{2}\left(\dfrac{1}{R+1}\right) & \text{for } N \geq 2 \end{cases}$$

$$R = 0, 1, 2, 3, \cdots, \tag{26}$$

where

$$R = \frac{r}{N} = \left[\frac{M-N}{N}\right] = \text{redundancy factor.}$$

For noninteger values of R, (24a) can easily be evaluated numerically, but expressing it in analytical form becomes cumbersome. The reason for the $N=1$ anomaly is outlined in Section II-C. The results of (26) are indicated in Fig. 7.

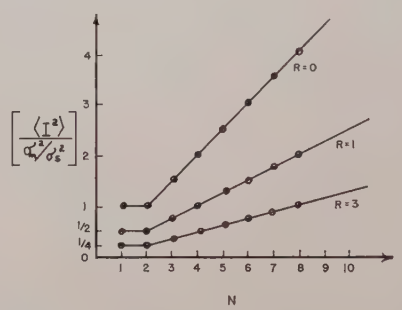


Fig. 7.—Variation of $\langle I^2 \rangle$ with N for $x(t)$ a square wave of period $2T_s$

A Technique of Linear System Identification Using Correlating Filters*

W. WAYNE LICHTENBERGER†, MEMBER, IRE

Summary—A technique for measuring the impulse response of linear processes while they are on line is described. Such an identification of process dynamics is necessary in process-adaptive control systems. A testing signal and correlating filter are employed after the manner of Turin. Such a procedure requires no multiplier, and the output of the filter is the impulse response as a continuous function of real time. To reduce accompanying output noise, the method of adding coherently the results of a number of tests made in succession is proposed. This idea is applied to the measurement of a member of an ensemble of slowly varying impulse responses. Optimum design of both the correlating filter and the necessary test signal is determined on the basis of minimum mean-square error of the resulting estimate. The optimization of the number of tests to be included in a measurement is described. The general results are applied to the case of a single, slowly time-varying process. In addition to optimum design, normalized curves showing the optimum number of tests for a particular mode of variation are included. A second application is made to the problem of measuring a member of an ensemble of fixed processes. The results of a digital computer simulation of this case are given.

Introduction

IT is important to identify transfer characteristics of a process in a process-adaptive control system. The extent to which this problem can be solved often dictates the ultimate design of the entire control system. When the process is linear, the desired information often sought is the impulse response. Although the impulse response conveys more information than is needed in some applications, we shall seek it because of its completeness and generality.

This paper describes a method of determining the impulse response of a linear process while it is on line. The method consists of injecting a special test signal into the input along with the regular actuating signal and then passing the output through a special filter. While the method is not basically superior to conventional correlating methods, it does offer certain advantages and may be looked upon as a better alternative for some applications.

* Received by the PGAC, December 15, 1960; revised manuscript received, March 15, 1961. Taken in part from a thesis submitted to the Graduate College of the University of Illinois, Urbana, in partial fulfillment of the requirements leading to the Ph.D. degree, January, 1961. The research reported in this paper was supported jointly by the Dept. of the Army (Signal Corps and Ordnance Corps), Dept. of the Navy (Office of Naval Res.), and the Dept. of the Air Force (Office of Scientific Res., Air Res. and Dev. Command) under Signal Corps Contract DA-36-039-SC-85122 with the Coordinated Science Lab., University of Illinois.

† Coordinated Science Lab. and Dept. of Elec. Engrg., University of Illinois, Urbana, Ill.

The Use of Statistical Methods for Impulse Response Determination

Well-established statistical methods [1] for determining the characteristics of processes unfortunately require the process to be investigated to be time-invariant. In practice, of course, processes requiring complex adaptive control systems are time-variable. If one applies techniques of conventional correlation to a situation which is in fact nonstationary, the results are inevitably in error. The magnitude of this error depends, however, on the rate of variation of the process parameters. The results will contain negligible error for *slowly varying* processes.

How slowly must a process vary in order to be considered slowly varying? The system function for a time-variable process is defined as

$$H(j\omega, t) = \int_{-\infty}^{\infty} h(\lambda, t)e^{-j\mu\lambda}d\lambda, \tag{1}$$

where $h(\lambda, t)$ is the response measured at time t due to a unit impulse applied λ seconds earlier. If we take the Fourier transform of $H(j\omega, t)$, we obtain a bifrequency response function

$$\mathcal{H}(j\omega, j\mu) = \int_{-\infty}^{\infty} H(j\omega, t)e^{-j\omega t}dt. \tag{2}$$

μ is the variable corresponding to the process variation, and ω is the variable corresponding to the output variation. We will assume that a process is slowly varying if the bandwidth in the ω domain for all values of t is ten times the bandwidth in the μ domain. That is to say, the impulse response decays to zero before the process parameters can vary significantly.

A direct approach to correlation schemes is carried out as follows: If the output $y(t)$ of a linear process (cf. Fig. 1) with impulse response $h(\lambda)$ is correlated with the

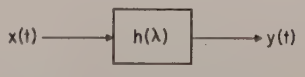


Fig. 1—A linear process.

input $x(t)$, the resulting cross-correlation is given by

$$R_{yx}(\tau) = \int_{-\infty}^{\infty} h(\lambda)R_{xx}(\tau - \lambda)d\lambda, \tag{3}$$

where τ is the time delay in the correlator. Eq. (3) has no particular value unless $R_{xx}(t)$ has a special form.

If $x(t)$ is such that

$$R_{xx}(t) = N_0\delta(t), \tag{4}$$

where $\delta(t)$ is the Dirac delta or unit impulse and N_0 a constant, then

$$R_{yx}(\tau) = \int_{-\infty}^{\infty} h(\lambda)N_0\delta(\tau - \lambda)d\lambda. \tag{5}$$

Thus,

$$R_{yx}(\tau) = N_0h(\tau). \tag{6}$$

The assumption of (4) is too great to render the results of (6) practical. $X(t)$ would have to have the spectral properties of white noise for this to be true.

Often [2] a low-level test signal with a flat power density spectrum is injected into the input along with the actuating signal. The process output is then correlated against this test signal.

The usual realization of the correlator is a time delay, multiplier, and integrator. The ensemble average is replaced with the time average under the assumption that all time series involved are ergodic. This is not strictly true with slowly-varying processes. Even if it were true, however, one would still have to average over-all time before the cross-correlation could be precisely determined. Since such an average must be performed in real time, a compromise must be made. Reducing the integration time permits the actuating signals (which for the purposes of identification must be regarded as noise) to appear at the correlator output.

Some disadvantages of such a correlator are, thus: 1) in order to make the output signal to noise ratio large, the integration time must be kept at some value T. The correlator can thus respond to changes in $h(\lambda)$ only after T seconds. 2) An ideal multiplier is required. 3) The output of the correlator is the impulse response evaluated at only one value of time. For each value of time desired a separate correlator must be used. This could lead to a great multiplicity of equipment if the process were such that a large number of sample points were necessary to determine completely the impulse response.

Correlation by Means of Linear Filters

Let us now compare the input-output relation of a fixed linear filter given by

$$y(t) = \int_{-\infty}^{\infty} h(\lambda)x(t - \lambda)d\lambda, \tag{7}$$

with the expression for cross-correlation given by

$$R_{hx}(t) = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} h(\lambda)x(\lambda + t)d\lambda. \tag{8}$$

An obvious difference in these equations is the $1/T$ factor of (8). This factor appears in the definition because the time-series treated up to now have been those of infinite energy but finite power. Such time-series are usually referred to [2] as being of infinite duration. If, however, $h(\lambda)$ or $x(t)$ is of *finite* duration,[1] then (8) must be redefined by

$$R_{hx}(t) = \int_{-\infty}^{\infty} h(\lambda)x(t + \lambda)d\lambda \tag{9}$$

in order to prevent $R_{hx}(t)$ from being trivially zero.

Further comparison of (7) and (9) can be made conveniently by considering their respective Fourier transforms, namely

$$Y(j\omega) = H(j\omega)X(j\omega)$$

and

$$S_{hx}(j\omega) = H^*(j\omega)X(j\omega), \tag{10}$$

where the capital letters denote the corresponding transforms; $S_{hx}(j\omega)$ is the transform of $R_{hx}(j\omega)$, and the asterisk denotes the complex conjugate. Eq. (10) shows that the output of a linear filter can be taken to be the cross-correlation of a function whose transform is the complex conjugate of the filter transfer function with the input function. Translating this statement into terminology of the time domain, the cross-correlation of two time functions (of finite duration) can be obtained by passing one of them through a linear filter whose impulse response is the time reverse of the other. The prospect of using this fact to accomplish the cross-correlation needed to determine impulse response is interesting. The use of an ideal multiplier is no longer necessary with this method, and the output of the filter is a continuous function of time.

Fig. 2(a) shows a filter in use as a correlator. A test signal $x(t)$ is passed into a process $h(\lambda)$ to be measured. The output is passed into the filter whose impulse response is $x(-t)$.[2] It is instructive to rearrange the blocks of Fig. 2(a) as shown in Fig. 2(b). Such an operation is permissible because of linearity. In Fig. 2(b) we see that $x(t)$ is first correlated with itself, and the resulting autocorrelation function is passed into the process undergoing measurement. The final output is thus the convolution of this autocorrelation function with the impulse response of the process. If the output is to be the impulse response of the process, it is clear that the autocorrelation function of the test signal must be an impulse.

There are two major limitations on this idea. First, $x(t)$ must have a flat spectrum of infinite bandwidth if $R_{xx}(t)$ is to be an impulse. Since this cannot be attained

[1] More properly if

$$\int_{-\infty}^{\infty} h^2(t)dt \quad \text{or} \quad \int_{-\infty}^{\infty} x^2(t)dt$$

is finite.

[2] We will not be concerned immediately with the condition of relizability, *i.e.*, that the impulse response must be zero for negative time. We may introduce time delays to avoid this.

in a practical situation, $R_{xx}(t)$ will have a nonzero width (roughly the inverse of the bandwidth), and the output will be distorted accordingly. Second, the measurements must be made in the presence of the normal actuating signals which for the purpose of measurement must be regarded as noise. For practical reasons the amplitude of the test signal must be small with respect to these signals so that the process output is not appreciably disturbed. We must rely upon the matching effect in the correlating filter to enhance the test signal to actuating signal power ratio.[3] Let us now ask questions of a more quantitative nature. How much error is caused by this scheme? Is there a better filter design than the matched filter? What is the best form of test signal?

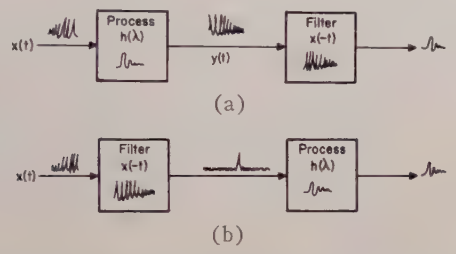


Fig. 2—(a) Measurement of impulse response by means of a correlating filter. (b) Alternate arrangement of the blocks of Fig. 2(a) for illustrative purposes.

Optimum Design for Single Measurements

The answers to these questions were supplied by Turin [4] in solving the problem illustrated in Fig. 3. A test signal $x(t)$ is added to noise $n_j(t)$ and passed into a member $h_i(\lambda)$ of an ensemble of processes. An "estimating" filter is placed at the output of the process, and the output of this filter is compared to the impulse response of the process. Turin solved for the particular estimating filter transfer function $H_{e\,\mathrm{opt}}(j\omega)$ and the particular test signal spectrum $X_{\mathrm{opt}}(j\omega)$ which minimized the expected mean-square error of the estimating filter output. This solution was made with respect to the ensemble of noises $n_j(t)$ and the ensemble of processes $h_i(\lambda)$. Two constraints placed upon the testing signal were that it have a constant energy and that it (and therefore the estimating filter) be band-limited to band B_1.

Although Turin's problem arose in the field of communications and radar, the applications to process identification for adaptive control are obvious. In our case the input "noise" is largely actuating signal. Of course *any* input to the process other than the test signal is considered to be noise. In the most general case, we have not only actuating signal present but also other signals of undisclosed origin. Some of the latter might even originate within the process. We shall, however, think of them as having been introduced at the input together with the actuating signal. All of these signals will be assumed to be mutually independent. Thus, their power spectra are additive to what we shall call the input noise power spectrum $S_n(j\omega)$.

[3] Filters of the type described here are known as "matched" filters [3] and are used especially in radar and communications to minimize the peak SNR.

A Technique for Reduction of Output Noise

In general, the "noise" power will be relatively greater than the test signal power. The output noise will accordingly be quite high. We, therefore, are confronted with the problem of increasing the energy of the test signal without increasing its average amplitude. This can be done only by increasing its duration. One way of increasing the effective duration of the test signal is to add the results of several measurements at each instant of time. This is a standard technique [5] for improving the SNR. Of course, resorting to such a scheme will lengthen the determination of impulse response by perhaps many times. The problem, however, is fundamental and also occurs in the standard techniques of correlation where it is necessary to increase the time constant of the integrators until the output SNR is tolerable. The coherent summation may be achieved in a variety of ways. A tapped delay line, a recirculating delay line, or other analog or digital equipment may be used for the purpose. In the following development the word "test" will be used to denote a single measurement like the measurement of Turin's. A "measurement" will refer to the integrated results of several tests. In addition processes undergoing measurement will be required to be fixed during a single test but will be allowed to vary from test to test. We thus have the following generalization of the problem of Turin.

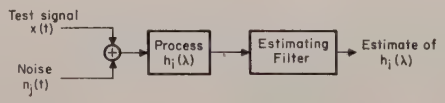


Fig. 3—Turin's identification problem.

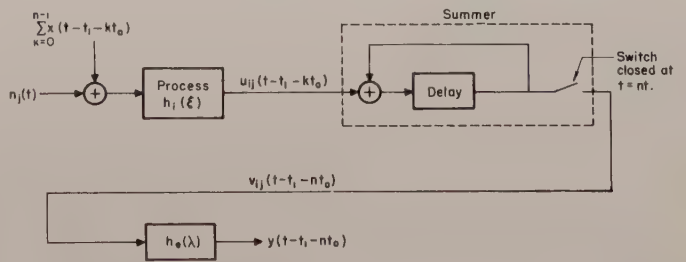


Fig. 4—Extension of Turin's problem to reduce the noise component in $y(t-t_1-nt_0)$.

A finite train of n test pulses is added to the normal input of a linear process (cf. Fig. 4). The output of the process during each test pulse is added to the sum of all previous tests in the measurement at each instant of time. This is accomplished by the use of a delay or storage element and an adder (not an integrator) as shown in Fig. 4. After the last test pulse, the result of the summation, still a function of time, is passed into an estima-

tion filter $h_e(\lambda)$. The expected mean-square error will be minimized by a choice of estimation filter, test signal, and the number of tests included in each summation.

Optimum Design for an n-Test Measurement

The expected mean-square error is given by

$$\epsilon = \left\langle\left\langle \frac{1}{T}\int_0^T [y_{ij}(t, t_1 + nt_0) - h_i(t, t_1 + nt_0)]^2 dt \right\rangle\right\rangle_{ij}, \quad (11)$$

where T is the maximum permissible duration of the process impulse responses in the ensemble being measured; t_1 is the time at which the measurement is started, t_0 the duration of each trial, and the $\langle\langle\ \ \rangle\rangle_{ij}$ signs denote averages over the ensembles of noises n_j and of processes h_i. We minimize the error first by varying the estimating filter.

It is shown in Appendix I that for an arbitrary testing signal the optimum design of the estimation filter is given by

$$H_e(j\omega) = \frac{\langle H_i \overline{H_i}^* \rangle_i X^*(j\omega)}{\langle |\overline{H}_i|^2\rangle_i |X(j\omega)|^2 + \frac{T}{n}\overline{\langle |H_i|^2\rangle_i} S_n(j\omega)}, \quad (12)$$

where for convenience $H_i = H_i(j\omega, t_1 + nt_0)$ is the transfer function of the sample process expressed also as a function of time;

$$\overline{H}_i = \frac{1}{n}\sum_{k=0}^{n-1} H_i(j\omega, t_1 + kt_0)$$

is the mean process transfer function over all the tests, $X(j\omega)$ is the transform of the test signal, $S_n(j\omega)$ is the noise power spectral density, and t_0 is the time between the start of each test. While it appears that $H_e(j\omega)$ is a function of the very process $H_i(j\omega)$ it is attempting to measure, $H_e(j\omega)$ is actually a function only of the *statistics* of the ensemble of H_i's, given by various ensemble averages. For time invariant processes, *i.e.*, for

$$H_i(j\omega, t) = H_i(j\omega), \quad (13)$$

(12) reduces to

$$H_e(j\omega) = \frac{X^*(j\omega)}{|X(j\omega)|^2 + \frac{T}{n} S_n(j\omega)}. \quad (14)$$

Except for the factor $1/n$ in the denominator, (14) is identical to Turin's (5).

The result of (12) is optimum for any test signal $x(t)$. Let us assume, then, that the estimation filter is now optimum and attempt to minimize further the mean-square error of (11) this time by choosing $x(t)$. We impose the constraints on $x(t)$ that its transform $X(j\omega)$ shall lie only in the band B_1 and that $x(t)$ shall have a fixed amount of average power W. Thus, we have

$$X(j\omega) = 0, \qquad \omega \text{ not in } B_1$$

and

$$\frac{1}{2\pi t_x}\int_{-\infty}^{\infty} |X(j\omega)|^2 d\omega = W, \quad (15)$$

where t_x is the duration of the test signal.

It is shown in Appendix II that the optimum test signal is given by

$$X(j\omega) = \begin{cases} G^{1/4}(j\omega, t_1, n, t_0) S_n^{1/4}(j\omega) \left\{ \left[\dfrac{|\langle H_i^* \overline{H}_i\rangle_i|^2}{n\lambda \langle |\overline{H}_i|^2\rangle_i} \right]^{1/2} \right. \\ \qquad \left. - \dfrac{T}{n} G^{1/2}(j\omega, t_1, n, t_0) S_n(j\omega) \right\}^{1/2} e^{-j\beta(\omega)}, \quad (16) \\ \qquad\qquad\qquad\qquad \omega \text{ in } B_2 \\ 0, \quad \omega \text{ in } \overline{B}_2, \end{cases}$$

where

$$G(j\omega, t_1, n, t_0) = \frac{\overline{\langle |H_i|^2\rangle_i}}{\langle |\overline{H}_i|^2\rangle_i} \quad (17)$$

and

$$\sqrt{\lambda} = \frac{\frac{\sqrt{n}}{2\pi}\int_{B_2} \frac{[|\langle H_i^* \overline{H}_i\rangle_i|^2 \overline{\langle |H_i|^2\rangle_i}]^{1/2} S_n^{1/2}(j\omega)}{\langle |\overline{H}_i|^2\rangle_i} d\omega}{nWt_x + \frac{T}{2\pi}\int_{B_2} G(j\omega, t_1, n, t_0) S_n(j\omega) d\omega}. \quad (18)$$

$\beta(\omega)$ is an arbitrary phase function to be elaborated on later, and B_2 is the sub-band of B_1 where the term in braces in (16) is real.

The meaning of the first factor $G^{1/4}(j\omega, t_1, n, t_0) S_n^{1/4}(j\omega)$ of (16) is that if the noise power at a given frequency is small, then little signal energy is needed at that frequency in order to determine $h_i(\lambda)$. On the other hand, the second factor (in the braces) states that if the noise power at a given frequency is large or if the average process power transmission is small, it is a waste of the finite amount of power available to put much, if any, at that frequency.

If (16) is substituted into (12), we have the expression for that particular estimation filter which will minimize the mean-square error, given n tests. We have

$$H_{e_{opt.}}(j\omega) = \begin{cases} \dfrac{\sqrt{n\lambda}\langle |H_i|^2\rangle_i G^{-1/4}(jw) S_n^{-1/4}(j\omega)}{[|\langle H_i^* \overline{H}_i\rangle_i|^2 \langle |\overline{H}_i|^2\rangle_i]^{1/2}} \left\{ \left[\dfrac{|\langle H_i^* \overline{H}_i\rangle_i|^2}{n\lambda \langle |\overline{H}_i|^2\rangle_i} \right]^{1/2} - \dfrac{T}{n} G^{1/2}(j\omega) S_n^{1/2}(j\omega) \right\}^{1/2} e^{+j\beta(\omega)}, & \omega \text{ in } B_2 \\ 0, \quad \omega \text{ in } \overline{B}_2. \end{cases} \quad (19)$$

By substituting the expressions for optimum estimating filter and optimum testing signal, given by (19) and (16), into (11), the mean-square error, we obtain

$$\epsilon = \frac{1}{2\pi T}\int_{\overline{B}_2} \langle |H_i|^2\rangle_i d\omega + \frac{1}{2\pi T}\int_{B_2}\left\{\frac{|\langle H_i^*\overline{H}_i\rangle_i|^2}{\langle|\overline{H}_i|^2\rangle_i} - \langle|H_i|^2\rangle_i\right\}d\omega + \frac{\left\{\frac{1}{2\pi}\int_{B_2}\frac{[|\langle H_i^*\overline{H}_i\rangle_i|^2\overline{\langle|H_i|^2\rangle_i}]^{1/2}}{\langle|\overline{H}_i|^2\rangle_i}S_n^{1/2}(j\omega)d\omega\right\}^2}{nWt_x + \frac{T}{2\pi}\int_{B_2}\frac{\overline{\langle|H_i|^2\rangle_i}}{\langle|\overline{H}_i|^2\rangle_i}S_n(j\omega)d\omega}. \quad (20)$$

The first term of (20) is error of omission, arising from the lack of test signal components in certain regions of the pass band of $h_i(\lambda)$. The presence of the second term of (20) may be attributed to the time-varying nature of the ensemble averages. If the processes are assumed to be fixed or if a single process is dealt with so that the ensemble averages are not necessary, this term drops out. The third term of (20) represents the components of error caused by the noise and the distortion of $h_i(\lambda)$ which occurs unavoidably when the noise is reduced. In addition, some time variation terms are included.

The phase function $\beta(\omega)$ appearing in (16) and (19) was left unspecified mainly since its choice did not directly affect the expression for minimum mean-square error given by (20). However, it is wrong to conclude that *any* choice of $\beta(\omega)$ is equally as good as another. It is, of course, the combination of the amplitude and phase spectra which determines the function in the time domain. Thus, in any situation where constraints are placed on the maximum amplitude (more properly, the peak power) of a function, the class of admissible phase functions is somewhat restricted.

Consider, for example, a signal having a flat amplitude spectrum over some band and a phase function which is linear with frequency. The inverse transform or time domain equivalent of such a signal is a narrow pulse, the shape of which depends on the way in which the amplitude spectrum cuts off at the edge of the band. If the peak power of such a pulse is limited, then so is the total energy in the pulse. It is helpful to look at the so-called "group delay," or $-d\beta/d\omega$. For the case of linear phase, $\beta(\omega) = -t_0\omega$. Thus, $-d\beta/d\omega = t_0$ and, hence, all components of energy are delayed by the same time, emerging together to give an energetic pulse. We need a group delay which allows a steady amount of energy to emerge per unit time. Fig. 5 presents the elements of this argument. In Fig. 6 is shown a more desirable group delay (assuming a constant energy density). It leads, of course, to a quadratic phase function which is the basis of the well-known linearly FM signal of modern high-resolution radars [7].

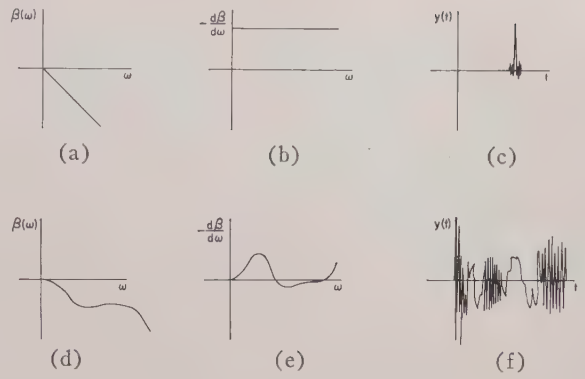


Fig. 5—Illustration of pulse dispersion with two types of phase characteristics in a network whose amplitude response is flat over a wide band. (a) Linearly decreasing phase. (b) Corresponding group delay. (c) Resulting impulse response. (d) Nonlinearly decreasing phase. (e) Corresponding group delay. (f) Resulting impulse response (not to same scale as above).

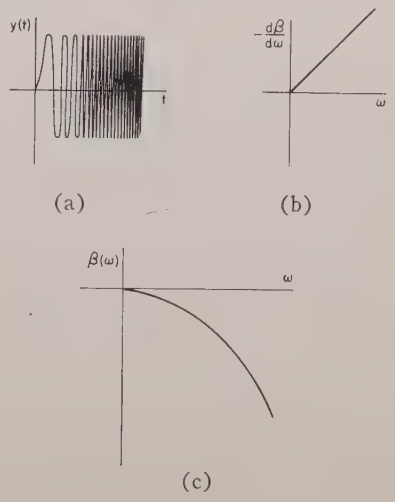


Fig. 6—Example of a desirable phase function. (a) A desirable impulse response. (b) The necessary group delay. (c) The required phase characteristic.

In fact the requirements on our phase functions and those found in the field of matched-filter radars are quite similar. We have a peak-power limited signal into which we are trying to crowd as much energy as possible. We do this not by increasing the amplitude but by increasing the duration of the pulse without changing its bandwidth. All this is accomplished by a judicious choice of the phase function. The interested reader is referred to the vast body of literature on the subject of matched-filter radars and related topics to which only a few references can be included [3].

Finally, turning our attention to (20), the expression for minimum mean-square error, we see that the error is yet a function of n, t_0, and t_x which may yet be varied in order to effect a further reduction of mean-square error. Of these, t_0 and t_x cannot be chosen independently. They are related by

$$t_0 = t_x + T, \quad (21)$$

i.e., the duration between the start of successive trials is

given by the sum of the maximum expected impulse response time and the duration of the test signal.

Thus, we are left only with the number of trials n and the time duration t_0 of each trial with which to work. The duration of a measurement—time required to sum the results of the n tests—is given by nt_0, the product of these variables. Further improvements in the measurement will come by optimizing this product. Such an optimization is a straightforward problem in the calculus.

Examples

Eqs. (16), (19), and (20) are complicated in appearance mainly because of their general application. Simplifications result when they are applied to specific problems. Of the general, slowly time-varying ensembles which we have considered, there are some interesting special cases. One such case is that of an ensemble of fixed processes (such as that of Turin) where the maximum number of trials permitted in the measurement scheme is determined by considerations other than minimum mean-square error. Another such case is that of a single, known impulse response which is slowly changing in time. The identification process must follow the course of the changes. It is this case which we shall first consider.

Suppose we have a single, known process which varies slowly with time. It is desired to follow the course of these variations so that the current state of the process will always be known. Under such conditions, the ensemble averages disappear from the previous equations, and the cancellations which can be made from the resulting equations reduce them into the forms given below. For the optimum estimation filter with arbitrary test signal, we have

$$H_e(j\omega) = \frac{H\overline{H^*}X^*(j\omega)}{|\overline{H}|^2 |X(j\omega)|^2 + \frac{T}{n}\overline{|H|^2}S_n(j\omega)}. \quad (22)$$

Before (22) can be reduced further we must assume a form of time variation of $h(\lambda, t_1)$. For this example, we will assume that the impulse response $h(\lambda, t_1)$ of the process is separable into a fixed part $h_f(\lambda)$ and a linearly time-varying gain a_0+at_1; *i.e.*, let

$$h(\lambda, t_1) = (a_0 + at_1)h_f(\lambda).$$

In the frequency domain we have

$$H(j\omega, t_1) = (a_0 + at_1)H_f(j\omega).$$

Substituting this into (22), we obtain

$$H_e(j\omega) = \frac{[a_0 + a(t_1 + nt_0)]\left[\frac{1}{n}\sum_{k=0}^{n-1} a_0 + a(t_1 + kt_0)\right]X^*(j\omega)}{\left[\frac{1}{n}\sum_{k=0}^{n-1} a_0 + a(t_1 + kt_0)\right]^2 |X(j\omega)|^2 + \frac{T}{n}\left[\frac{1}{n}\sum_{k=0}^{n-1}(a_0 + a(t_1 + kt_0))^2\right]S_n(j\omega)}. \quad (23)$$

Thus, we have a case similar to that of (14) but with time-variable gains on each term.

Upon dropping the ensemble averages and simplifying, (16), the optimum test signal, becomes

$$X(j\omega) = \begin{cases} G^{1/4}(j\omega)S_n^{1/4}(j\omega)\left[\frac{1}{\sqrt{n\lambda}}|H(j\omega, t_1 + nt_0)| - \frac{T}{n}G^{1/2}(j\omega)S_n^{1/2}(j\omega)\right]^{1/2} e^{-j\beta(\omega)}, & \omega \text{ in } B_2 \\ 0, \quad \omega \text{ in } \overline{B_2} \end{cases} \quad (24)$$

where

$$G(j\omega) = \frac{\overline{|H|^2}}{|\overline{H}|^2},$$

$$\sqrt{n\lambda} = \frac{\frac{1}{2\pi}\int_{B_2}[G(j\omega)|H(j\omega, t_1 + nt_0)|^2 S_n(j\omega)]^{1/2}d\omega}{Wt_x + \frac{T}{n}\frac{1}{2\pi}\int_{B_2} G(j\omega)S_n(j\omega)d\omega},$$

and B_2 is defined as that range of W where

$$\frac{1}{\sqrt{n\lambda}}|H(j\omega, t_1 + nt_0)| - \frac{T}{n}G^{1/2}(j\omega)S_n^{1/2}(j\omega) \geq 0.$$

For our example, $G(j\omega)$ becomes

$$G(j\omega) = \frac{\frac{1}{n}\sum_{k=0}^{n-1}[a_0 + a(t_1 + kt_0)]^2}{\left[\frac{1}{n}\sum_{k=0}^{n-1} a_0 + a(t_1 + kt_0)\right]^2} = \frac{a_0^2 + (n-1)a_0at_0 + \frac{(2n-1)(n-1)}{6}a^2t_0^2}{a_0^2 + (n-1)a_0at_0 + \frac{(n-1)^2}{4}a^2t_0^2}. \quad (25)$$

Divide both numerator and denominator of (25) by a_0^2 and let

$$\tau = \frac{at_0}{a_0}.$$

τ is proportional both to the normalized slope of the amplitude function and the time duration of a single trial. It is, thus, a measure of the amount of variation

which occurs during one trial. We have, then, in terms of τ

$$G = G(n, \tau) = \frac{1 + (n-1)\tau + \dfrac{(2n-1)(n-1)}{6}\tau^2}{1 + (n-1)\tau + \dfrac{(n-1)^2}{4}\tau^2}, \quad \begin{cases} \tau \geq 0 \\ n \geq 2 \end{cases} \tag{26}$$

Applying similar simplifications to the expression for mean-square error, (20) becomes

$$\epsilon = \frac{1}{2\pi T}\int_{\overline{B_2}} |\, H(j\omega, t_1 + nt_0)\,|^2 d\omega + \frac{(a_0 + ant_0)^2 G(n,\tau)\left[\dfrac{1}{2\pi}\displaystyle\int_{B_2} |\, H_f(j\omega)\,|\, S_n^{1/2}(j\omega)d\omega\right]^2}{nWt_x + TG(n,\tau)\dfrac{1}{2\pi}\displaystyle\int_{B_2} S_n(j\omega)d\omega}. \tag{27}$$

We will assume that $|H(j\omega, t_1+nt_0)|^2$ is sufficiently small in the region of B_2 such that the first term of (27) may be disregarded. We will be concerned with the second term ϵ_n which contains the contributions of the noise and smear components of the error.

Notice that

$$\frac{T}{2\pi}\int_{B_2} S_n(j\omega)d\omega$$

is the noise energy N per trial and that

$$\left[\frac{1}{2\pi}\int_{B_2} |\, H_f(j\omega)\,|\, S_n^{1/2}(j\omega)d\omega\right]^2$$

is related to the output noise power. Call this expression P_n. Making these substitutions and dividing numerator and denominator by NG, we have

$$\epsilon_n = a_0^2 \frac{P_n}{N}\frac{(1+n\tau)^2}{1 + \dfrac{n}{G}\left(\dfrac{Wt_x}{N}\right)} \tag{28}$$

where $G(n, \tau)$ is given by (26). We see that ϵ_n is a function of n, τ, and Wt_x/N. Wt_x/N will be recognized as the input test signal to noise (actuating signal) energy ratio.

A minimization program was prepared for Illiac, the University of Illinois digital computer. The function minimized was the normalized version of (28), namely,

$$\epsilon_N = \frac{(1+n\tau)^2}{1 + n\left(\dfrac{Wt_x}{N}\right)\dfrac{1}{G}}. \tag{29}$$

Plots of ϵ_N vs n with τ as a parameter are presented in Figs. 7–10 (next page) for four values of Wt_x/N. It is apparent from the figures that for a given value of Wt_x/N, as τ decreases (hence the rate of variation), we may integrate more trials before the point of diminishing return is reached. Hence, the minimum ϵ_N is smaller for each value of τ. Further, as Wt_x/N increases we may tolerate, in general, greater values of τ. This is intuitively obvious. The quantitative information conveyed by the figures is made more interesting if we note that the point $\epsilon_N = 1$ is that error which would be given by Turin's filter. Thus, we may do either better or worse than Turin, depending on the normalized rate of variation of the process and the input test to actuating signal power ratio.

Another case of special interest is the measurement of one of an ensemble of fixed processes where the number of trials n is limited by external reasons. Such a limitation might be imposed in order to effect a compromise between a low value of mean-square error and having the information available as soon as possible. For this case, (12) becomes

$$H_e(j\omega) = \frac{X^*(j\omega)}{|\, X(j\omega)\,|^2 + \dfrac{T}{n}S_n(j\omega)} \tag{30}$$

as has already been shown.

For the optimum test signal, (16) becomes

$$X_{\text{opt}}(j\omega) = \begin{cases} S_n^{1/4}(j\omega)\left[\dfrac{\langle |\, H_i(j\omega)\,|^2\rangle_i^{1/2}}{\sqrt{n\lambda}} - \dfrac{T}{n}S_n^{1/2}(j\omega)\right]^{1/2} e^{-j\beta(\omega)}, & \omega \text{ in } B_2 \\ 0, & \omega \text{ in } \overline{B_2} \end{cases} \tag{31}$$

since $G(j\omega) = 1$ for all fixed processes. Also, we have

$$\sqrt{n\lambda} = \frac{\dfrac{n}{2\pi}\displaystyle\int_{B_2}[\langle |\, H_i(j\omega)\,|^2\rangle_i S_n(j\omega)]^{1/2}d\omega}{nWt_x + \dfrac{T}{2\pi}\displaystyle\int_{B_2} S_n(j\omega)d\omega}$$

for this case. Thus, when the optimum test signal is used,

$$H_{e_{\text{opt}}}(j\omega) = \begin{cases} \dfrac{\sqrt{n\lambda}\, S_n^{-1/4}(j\omega)}{\langle |\, H_i(j\omega)\,|^2\rangle_i^{1/2}}\left[\dfrac{\langle |\, H_i(j\omega)\,|^2\rangle_i^{1/2}}{\sqrt{n\lambda}} - \dfrac{T}{n}S_n^{1/2}(j\omega)\right]^{1/2} e^{+j\beta(\omega)}, & \omega \text{ in } B_2 \\ 0, & \omega \text{ in } \overline{B_2}. \end{cases} \tag{32}$$

It should be noted that if the input noise is white and if there is no *a priori* knowledge of the statistics of the ensemble, then both $S_n(j\omega)$ and $\langle |\, H_i(j\omega)\,|^2\rangle_i$ are constants, and $H_{e_{\text{opt}}}(j\omega)$ is proportional to $X^*(j\omega)$. Thus, the optimum estimator is a matched filter. Furthermore, as shown by (31), the test signal has a flat amplitude spectrum with an arbitrary phase function. This agrees with the intuitive ideas expressed earlier.

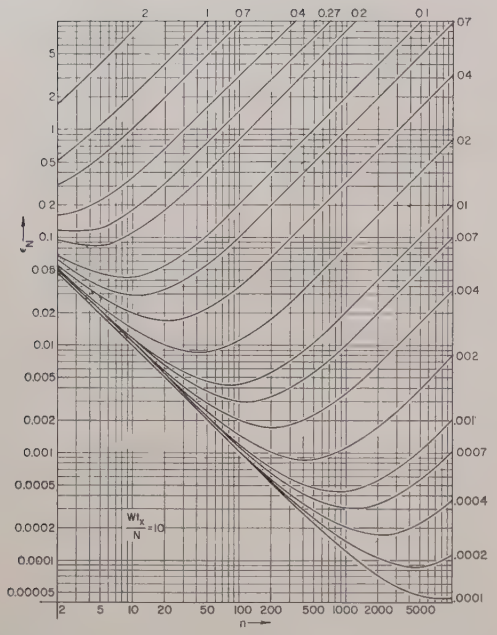


Fig. 7—Normalized error vs number of tests included in one measurement for 19 values of τ. Test signal to actuating signal energy ratio is 10.0.

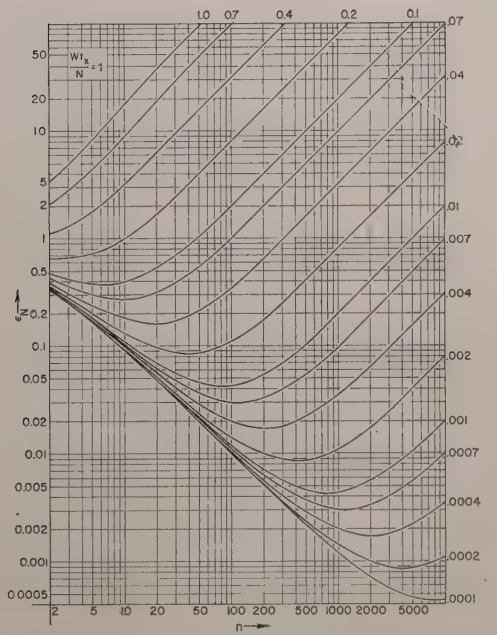


Fig. 8—Normalized error vs number of tests included in one measurement for 17 values of τ. Test signal to actuating signal energy ratio is 1.0.

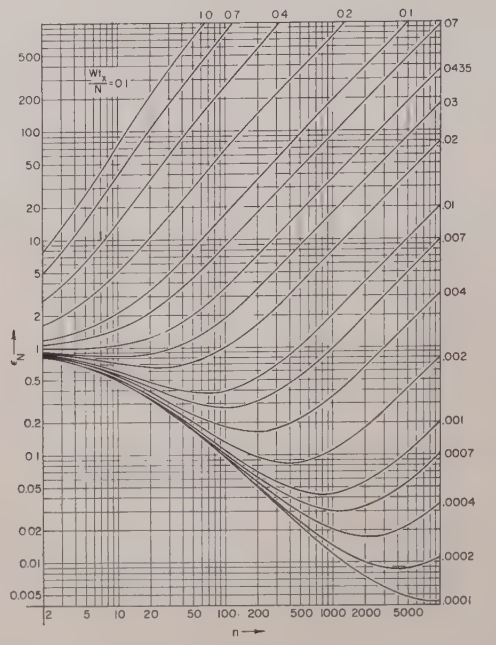


Fig. 9—Normalized error vs number of tests included in one measurement for 18 values of τ. Test signal to actuating signal energy ratio is 0.1.

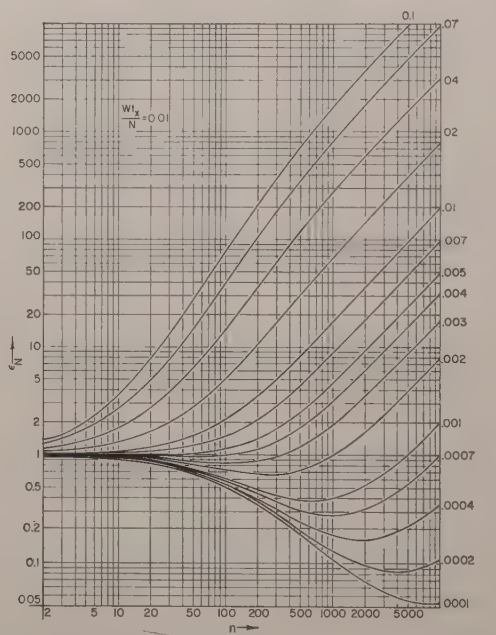


Fig. 10—Normalized error vs number of tests included in one measurement for 15 values of τ. Test signal to actuating signal energy ratio is 0.01.

There are other circumstances under which (30) leads to a matched filter. If

$$|X(j\omega)|^2 \gg \frac{T}{n} S_n(j\omega),$$

then

$$H_e(j\omega) = \frac{1}{X(j\omega)} \cdot \tag{33}$$

Although (33) is the expression for an inverse filter, we are free to choose *any* test signal since the resulting error is zero. Thus, there is no uniquely optimum test signal.

Obviously such a test signal should include components at all frequencies over the pass band of $H_s(j\omega)$. If this were not so, the inverse filter would have to have infinite transfer characteristics at all frequencies where the testing signal has no components. In fact, it is easy to see that a testing signal with a flat amplitude spectrum over the pass band of the process would be desirable to use since the corresponding inverse filter would be an all-pass network with the negative phase function. Notice, however, that for the case of an all-pass function and only for this case, the inverse function and the conjugate function are proportional. Thus, if we choose a test signal with a flat amplitude spectrum, we again have a matched filter for the optimum estimator.

Finally, when $(T/n)S_n(j\omega) \gg |X(j\omega)|^2$ we have

$$H_e(j\omega) = \frac{n}{T} \frac{X^*(j\omega)}{S_n(j\omega)} \cdot \tag{34}$$

If the input signal is white over the pass band of the process, then $H_e(j\omega)$ is matched to $X(j\omega)$ regardless of its form.

In summary, for all possible input SNR if the input noise is white over the pass band of the process and if no statistical knowledge of the ensemble of process is available *a priori*, then a test signal with a flat amplitude spectrum (over the pass band of the process) and its corresponding matched filter are optimum in a minimum mean-square sense.

The above results were applied in a simulation on Illiac to a process in the presence of actuating signals. The matched filter used in the simulation is described in Fig. 11. Both pole-zero plots and the impulse response are given. Theoretically [8] an infinitely long linear array of all-pass quadrupoles has for its impulse response a pure time delay. This is because its amplitude spectrum is flat and its phase spectrum is a linear function of frequency. By leaving out some of the quadrupoles, this linear behavior of the phase is disturbed so that the derivative of the phase function—the so-called "group delay"—is no longer a constant but is now a function of frequency. The impulse response is therefore no longer pure delay. A judicious choice of the pole-zero locations

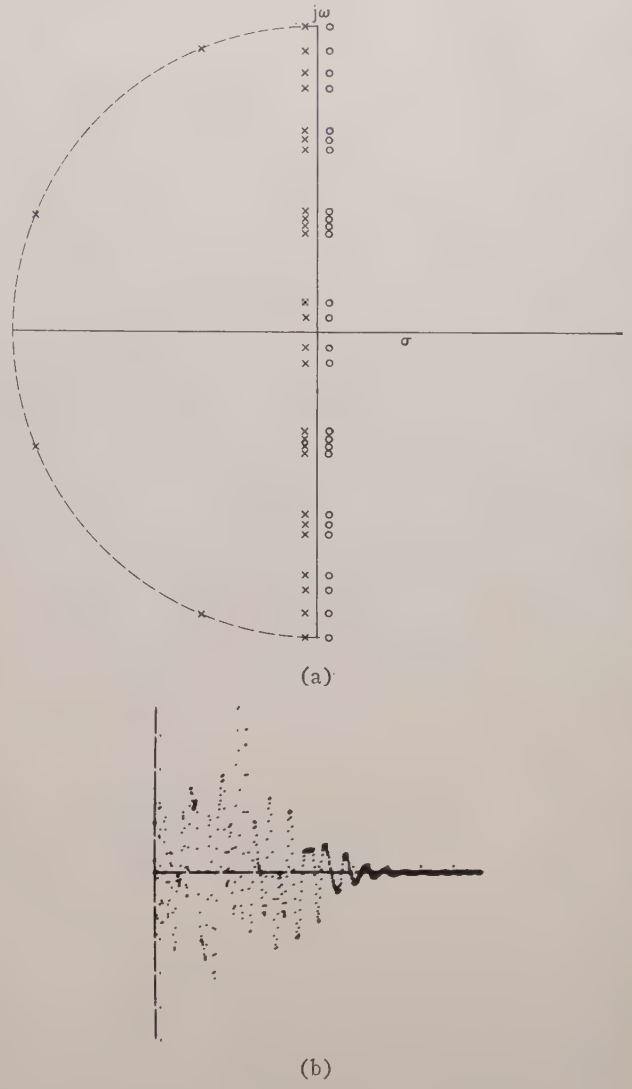


Fig. 11—(a) Pole-zero plot of a matched filter used as correlator in simulations. (b) Impulse response of the above filter. The time reverse of this function is the test signal.

enables the energy to be distributed more evenly in time. The Butterworth filter used in conjunction with the linear array helps to prevent a burst of energy at $t=0$.

Results of the simulation for two cases are shown in Figs. 12 and 13. In Fig. 12 no actuating signal is present. This is the ideal situation from the standpoint of measurement. In the figure, (a) is the input to the process. This input consists of the test signal only for this case. (b) is the resulting output of the process. The process for the example was second-order with a damping factor of 0.6. (c) is the output of the matched filter. Note the time delay which is necessary because of realizability. (d) is the impulse response to be determined. This is included for purposes of comparison. (e) is an expanded version of the interesting portion of (c), showing the resulting estimate of the process impulse response.

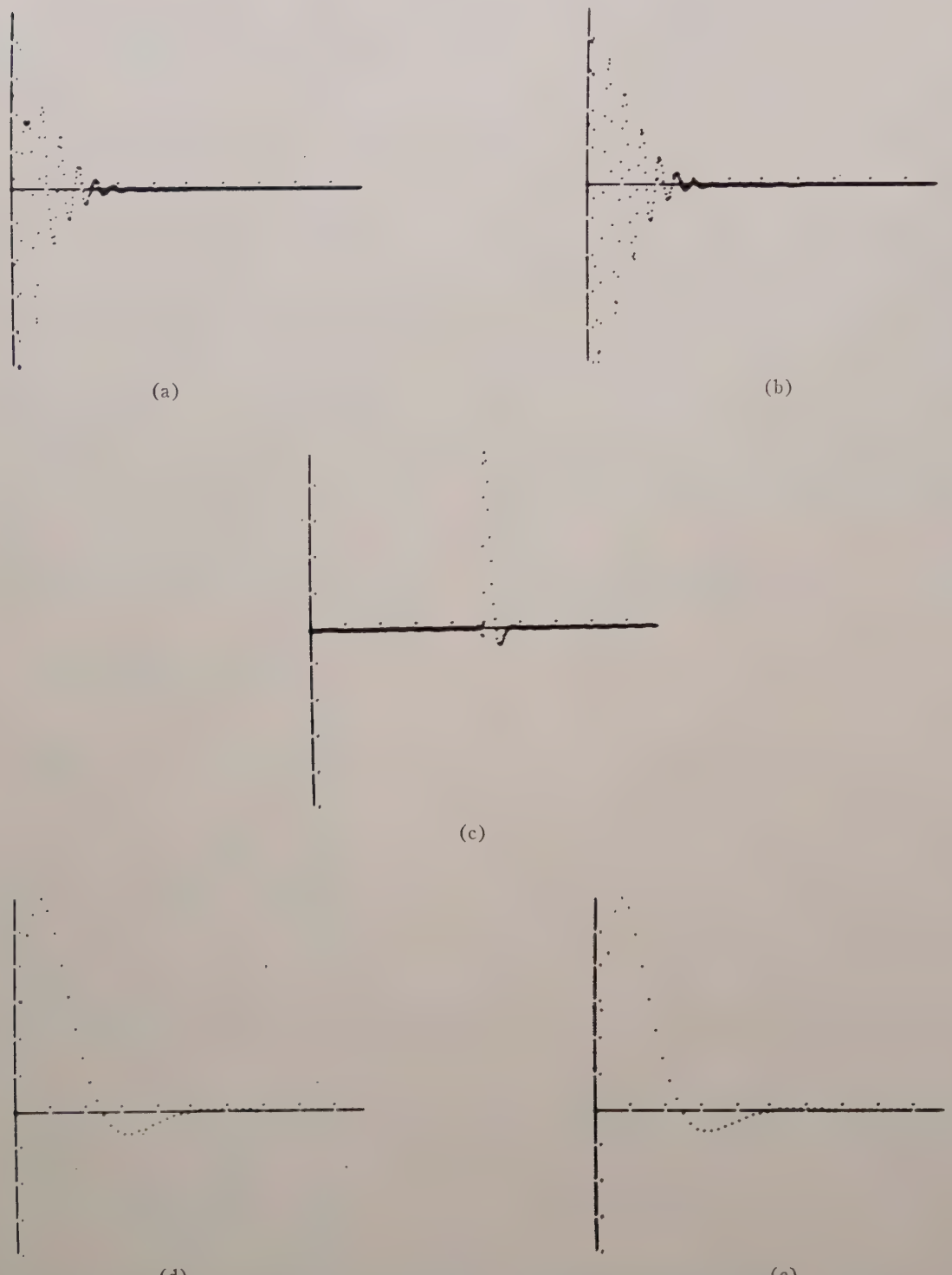


Fig. 12—Normalized results of a simulation of the measurement system in the absence of actuating signals. (a) Process input (test signal only for this case). (b) Process output and input to the correlating filter. (c) Correlating filter output. (d) Actual process impulse response. (e) Expanded version of (c). Measured impulse response.

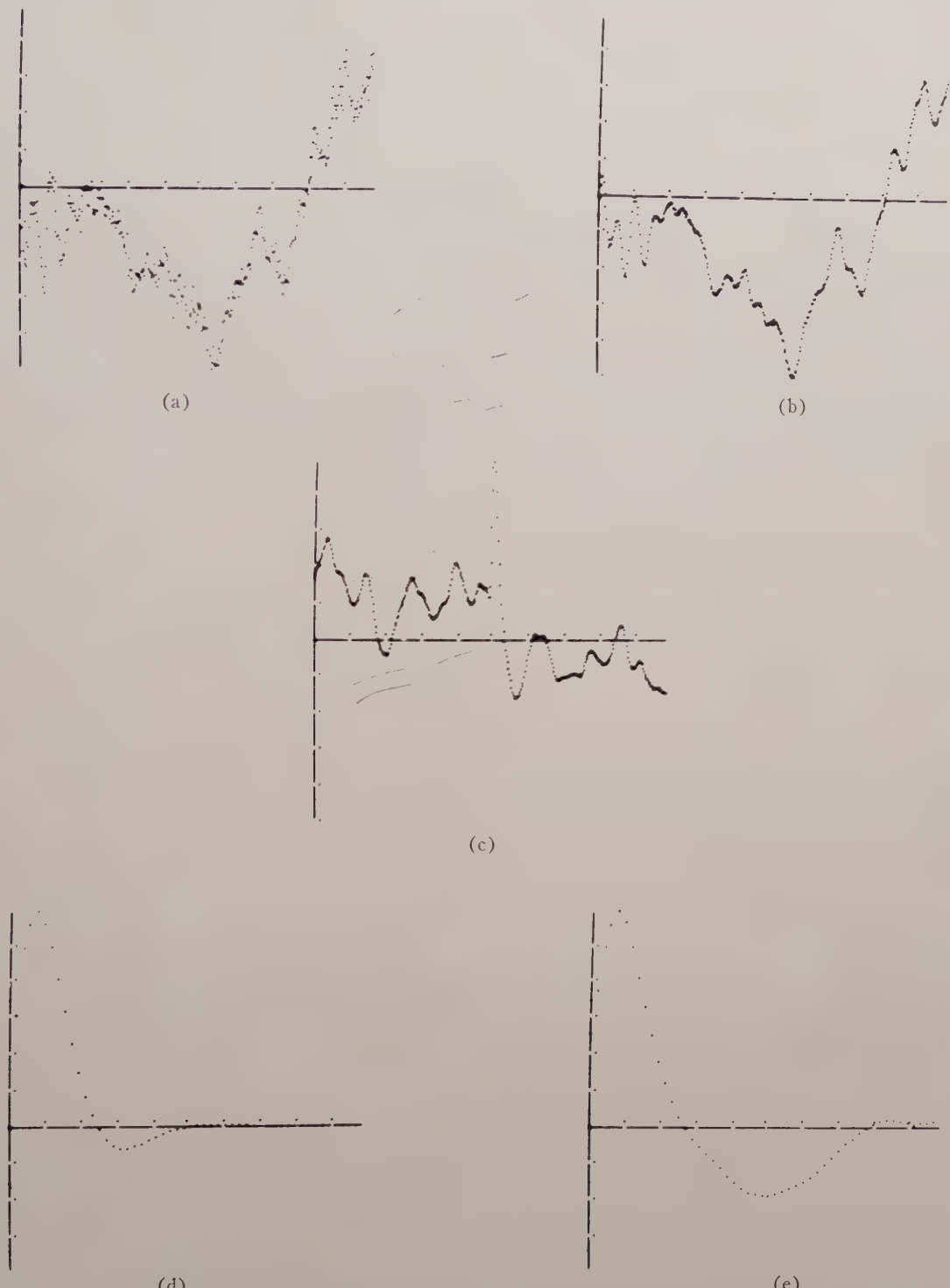


Fig. 13—Normalized results of a simulation of the measurement system with a test signal to actuating signal energy ratio of 0.73. (a) Process input (test signal plus actuating signal). (b) Process output and input to the correlating filter. (c) Correlating filter output. (d) Actual process impulse response. (e) Expanded version of (c). Measured impulse response.

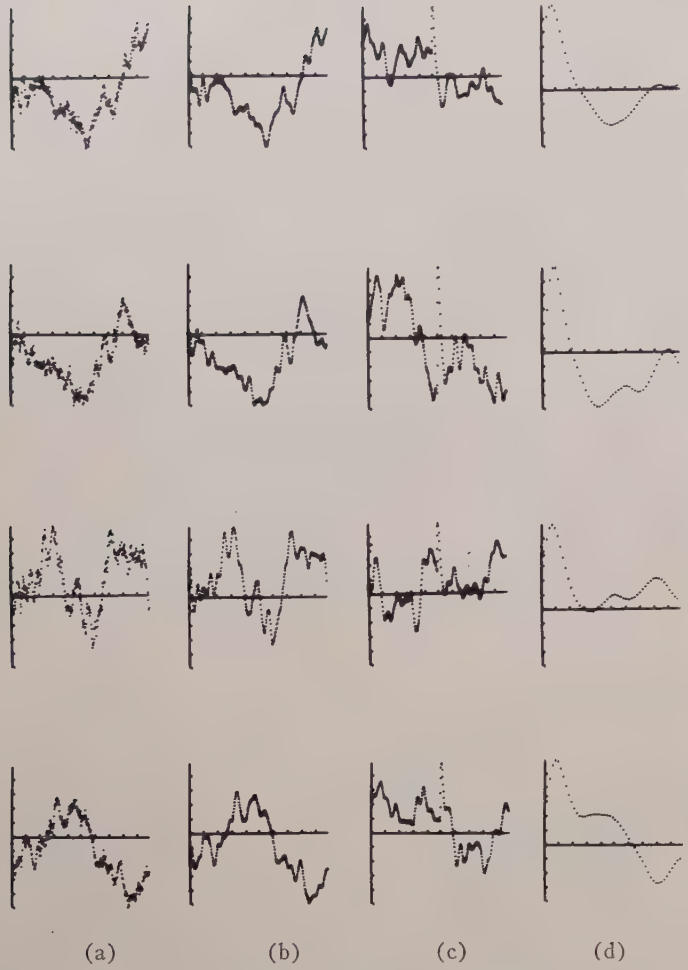


Fig. 14—Results of some tests included in a single measurement. (c) and (d) give the filter output for each test. This is only for convenience. Actually the signals of (b) could be added together before introduction to the filter. (a) Process inputs. (b) Process outputs. (c) Filter outputs. (d) Estimated impulse responses.

Fig. 13 gives the results of a similar test, this time with actuating signal present. This actuating signal is a random signal with a power spectrum which is flat over the pass band of the process. It was obtained by smoothing a table of random numbers. For this case the test signal to actuating signal energy ratio was 0.73. Notice the resulting disturbance to the estimate (e).

In Fig. 14, are given the results of some of the individual tests made. The test signal to actuating signal energy ratio was 0.1 for these tests. The samples were chosen to illustrate the diversity of the individual estimates. There is actually a coherent component present in each estimate which agrees with Fig. 12(e). The results of the first 25 runs were summed. This improved the effective test signal to actuating signal ratio to 0.5. The result of the summation is displayed in Fig. 15. The error present could have been reduced further, of course, by summing the results of more tests at the expense of time. Further, the process would have had to remain fixed for such an additional duration.

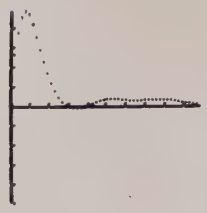


Fig. 15—Correlating filter output when the input is the integrated result of 25 successive tests. The effective improvement of test signal to actuating signal energy ratio is a factor of five.

Conclusions

In the determination of impulse response of on-line linear processes, correlating filters may in many cases prove useful. Their use becomes advantageous when a large number of sample points of the impulse response is desired, since the output of such filters is a continuous function of time. In these cases, the amount of equipment required may be less than that necessary for other correlation schemes. Other advantages of the filters are that time delays and multipliers are not required.

As with any short computation-time procedure, a single measurement will exhibit a large variance. This variance can be reduced by summing the results coherently with an accompanying expenditure of time and equipment. This is unfortunate because one would like to have the results available as soon as possible from the point of view of control system performance and because even for slowly varying processes, error builds up seriously for sufficiently long measurements. This is not basically different, however, from conventional correlation methods. Only the parameters of a particular situation may determine in general which procedure is more useful.

The most serious problem with correlating filters is the determination of a signal imbedded in noise. Much is known about this subject, and many of the techniques can be applied to this situation. Only the coherent integration approach has been discussed in this paper. Undoubtedly, however, there exist better approaches in given situations and perhaps in general.

Appendix I

Consider a finite train of test signals $x(t-t_1-kt_0)$ being fed into a member of an ensemble of slowly varying linear processes $h_i(\xi, t)$ along with the actuating signal $n_j(t)$ (cf. Fig. 4). For simplicity let us translate the time axis so that we will actually consider the input $x(t)+n_j(t+t_1+kt_0)$. t_1 is the time the measurement is started, and t_0 is the time between the start of each test. Define $n_{j+k}(t)=n_j(t+t_1+kt_0)$. The output of the process is then given by

$$u_{ij}(t, t_1 + kt_0) = \int_{-\infty}^{\infty} h_i(\xi, t_1 + kt_0)[x(t - \xi) + n_{j+k}(t - \xi)]d\xi. \quad (35)$$

These u_{ij} are fed into the summing device, and *after* all n of the tests has been made, *i.e.*, after nt_0 seconds the summer gives for its output,

$$v_{ij}(t,\ t_1\ +\ nt_0) = \frac{1}{n}\sum_{k=0}^{n-1} u_{ij}(t, t_1 + kt_0). \quad (36)$$

The output of the summer is passed into an estimation filter $h_e(\lambda)$, whose output is in turn given by

$$y_{ij}(t, t_1 + nt_0) = \int_{-\infty}^{\infty} h_e(\lambda)v_{ij}(t - \lambda, t_1 + (n-1)t_0)d\lambda. \quad (37)$$

$x(t)$, what choice of $h_e(\lambda)$ will minimize ϵ of (39)? We proceed by methods of the variational calculus. We have

$$\delta\epsilon = \left\langle\left\langle \frac{2}{T}\int_0^T [y_{ij}(t, t_1 + nt_0) - h_i(t, t_1 + nt_0)]\delta y_{ij}(t, t_1 + nt_0)dt \right\rangle\right\rangle_{ij} = 0, \quad (40)$$

where the notation $\delta\epsilon$ stands for "the variation of ϵ." Substituting (38) into (40) we get

$$0 = \left\langle\left\langle \int_0^T \left\{\frac{1}{n^2}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\int\int\int\int_{-\infty}^{\infty} h_e(\lambda)h_i(\xi, t_1 + kt_0)\delta h_e(\mu)h_i(\nu, t_1 + lt_0)[x(t - \lambda - \xi)x(t - \mu - \nu) + x(t - \lambda - \xi)n_{j+l}(t - \mu - \nu) + x(t - \mu - \nu)n_{j+k}(t - \lambda - \xi) + n_{j+k}(t - \lambda - \xi)n_{j+l}(t - \mu - \nu)]d\xi d\lambda d\nu d\mu - h_i(t, t_1 + nt_0)\frac{1}{n}\sum_{l=0}^{n-1}\int\int_{-\infty}^{\infty}\delta h_e(\mu)h_i(\nu, t_1 + lt_0)[x(t - \mu - \nu) + n_{j+l}(t - \mu - \nu)]d\nu d\lambda\right\} dt\right\rangle\right\rangle_{ij}. \quad (41)$$

Substituting (35) and (36) into (37) we have

$$y_{ij}(t, t_1 + nt_0) = \frac{1}{n}\sum_{k=0}^{n-1}\int\int_{-\infty}^{\infty} h_e(\lambda)h_i(\xi, t_1 + kt_0) \cdot [x(t - \lambda - \xi) + n_{j+k}(t - \lambda - \xi)]d\xi d\lambda. \quad (38)$$

Assuming the noise term is a member of a stationary random process with zero mean, we may perform the indicated ensemble average over the ensemble of noises n_j. This gives

$$0 = \left\langle \int_0^T \left\{\frac{1}{n^2}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\int\int\int\int_{-\infty}^{\infty}\delta h_e(\mu)h_e(\lambda)h_i(\xi, t_1 + kt_0)h_i(\nu, t_1 + lt_0)x(t - \lambda - \xi)x(t - \mu - \nu)d\nu d\xi d\lambda d\mu + \frac{1}{n^2}\sum_{k=0}^{n-1}\int\int\int\int_{-\infty}^{\infty}\delta h_e(\mu)h_e(\lambda)h_i(\xi, t_1 + kt_0)h_i(\nu, t_1 + kt_0)R_{nn}(\lambda + \xi - \mu - \nu)d\nu d\xi d\lambda d\mu - \frac{1}{n}\sum_{l=0}^{n-1}\int\int_{-\infty}^{\infty}\delta h_e(\mu)h_i(\nu, t_1 + lt_0)h_i(t, t_1 + nt_0)x(t - \mu - \nu)d\nu d\mu\right\} dt\right\rangle_i, \quad (42)$$

From (11) the average mean-square error ϵ is given by

$$\epsilon = \left\langle\left\langle \frac{1}{T}\int_0^T [y_{ij}(t, t_1 + nt_0) - h_i(t, t_1 + nt_0)]^2 dt \right\rangle\right\rangle_{ij}. \quad (39)$$

The problem is: given an arbitrary number n of trials, the statistics of the noise n_j, and an arbitrary test signal

where $R_{nn}(t)$ is the autocorrelation function of the noise. We interchange the order of integration, and extend the limits of integration with respect to t if we recognize that $h_i(t)$ is zero outside the range $0 \leq t < T$ and if we assume that the width of the central pulse of the autocorrelation function of $x(t)$ is small compared to T. We thus have to a good approximation

$$0 \approx \int_{-\infty}^{\infty}\delta h_e(\mu)\left\langle \frac{1}{n^2}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\int\int\int\int_{-\infty}^{\infty} h_e(\lambda)h_i(\xi, t_1 + kt_0)h_i(\nu, t_1 + lt_0)x(t - \lambda - \xi)x(t - \mu - \nu)dt d\nu d\xi d\lambda + \frac{T}{n}\frac{1}{n}\sum_{k=0}^{n-1}\int\int\int_{-\infty}^{\infty} h_e(\lambda)h_i(\xi, t_1 + kt_0)h_i(\nu, t_1 + kt_0)R_{nn}(\lambda + \xi - \mu - \nu)d\nu d\xi d\lambda - \frac{1}{n}\sum_{l=0}^{n-1}\int\int_{-\infty}^{\infty} h_i(\nu, t_1 + lt_0)h_i(t, t_1 + nt_0)x(t - \mu - \nu)dt d\nu\right\rangle_i d\mu. \quad (43)$$

We may avoid great complication by ignoring the requirement of physical realizability for $h_e(t)$, *i.e.*, that

$$h_e(t) = 0, t < 0.$$

Thus, we may say that $\delta h_e(\mu)$ is arbitrary for all values of μ. Eq. (43) cannot, therefore, be satisfied unless the bracketed term is zero. Equating this term to zero and taking the Fourier transform of the result, we have

$$0 = \left\langle \frac{1}{n^2} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} H_e(j\omega) H_i(j\omega, t_1 + kt_0) \cdot H_i^*(j\omega, t_1 + lt_0) \mid X(j\omega) \mid^2 + \frac{T}{n} \frac{1}{n} \sum_{k=0}^{n-1} H_e(j\omega) \mid H_i(j\omega, t_1 + kt_0) \mid^2 S_n(j\omega) - \frac{1}{n} \sum_{k=0}^{n-1} H_i^*(j\omega, t_1 + kt_0) H(j\omega, t_1 + nt_0) X^*(j\omega) \right\rangle_i. \quad (44)$$

Simplifying, taking ensemble averages of individual terms, and solving the result for $H_e(j\omega)$, we obtain

$$H_{e_{opt}}(j\omega) = \frac{\left\langle H_i(j\omega, t_1 + nt_0) \frac{1}{n} \sum_{k=0}^{n-1} H_i^*(j\omega, t_1 + kt_0) \right\rangle_i X^*(j\omega)}{\left\langle \left| \frac{1}{n} \sum_{k=0}^{n-1} H_i(j\omega, t_1 + kt_0) \right|^2 \right\rangle_i \mid X(j\omega) \mid^2 + \frac{T}{n} \left\langle \frac{1}{n} \sum_{k=0}^{n-1} \mid H_i(j\omega, t_1 + kt_0) \mid^2 \right\rangle_i S_n(j\omega)} \quad (45)$$

Appendix II

Eq. (45) gives the transfer function for the optimum estimation filter for an arbitrary number of trials when an arbitrary test signal is used. We shall assume now that such an optimum filter is used. The problem will be to further minimize the average mean-square error by choosing this time the optimum testing signal. The number of measurements will still be arbitrary.

Two constraints on $x(t)$ are that it shall be band-limited to band B_1 and that it shall have constant average power. Thus, we have

$$X(j\omega) = 0 \qquad \text{for } \omega \text{ not in } B_1 \quad (46)$$

and

$$\frac{1}{2\pi t_x} \int_{B_1} \mid X(j\omega) \mid^2 d\omega = W, \quad \text{where } t_x \text{ is the duration of } x(t). \quad (47)$$

Before we may minimize ϵ by varying $x(t)$, we must obtain an expression for it when $h_e(\lambda)$ is used. Squaring (39), we have

$$\epsilon = \left\langle \left\langle \frac{1}{T} \int_0^T [y_{ij}^2(t, t_1 + nt_0) - 2y_{ij}(t, t_1 + nt_0) h_i(t, t_1 + nt_0) + h_i^2(t, t_1 + nt_0)] dt \right\rangle \right\rangle_{ij}. \quad (48)$$

We substitute into this (38) to get

$$\epsilon = \left\langle \left\langle \frac{1}{T} \int_0^T \left[\frac{1}{n^2} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} \int\int\int\int_{-\infty}^{\infty} h_e(\mu) h_e(\lambda) h_i(\xi, t_1 + kt_0) h_i(\nu, t_1 + lt_0) [x(t - \lambda - \xi) x(t - \mu - \nu) + x(t - \lambda - \xi) n_{j+l}(t - \mu - \nu) + x(t - \mu - \nu) n_{j+k}(t - \lambda - \xi) + n_{j+k}(t - \lambda - \xi) n_{j+l}(t - \mu - \nu)] d\nu d\xi d\lambda d\mu - 2 \frac{1}{n} \sum_{l=0}^{n-1} \int\int_{-\infty}^{\infty} h_e(\mu) h_i(\nu, t_1 + lt_0) h_i(t, t_1 + nt_0) [x(t - \mu - \nu) + n_{j+l}(t - \mu - \nu)] d\nu d\mu + h_i^2(t, t_1 + nt_0) \right\} dt \right\rangle \right\rangle_{ij}. \quad (49)$$

Eq. (45) is the solution of the problem.

It is unfortunate that physical realizability cannot be easily included in the work leading to (45). The equation itself is the counterpart of the solution of Wiener's optimum prediction and filtering problem for the unrealizable case.

We now take the ensemble average over the ensemble of noises n_j, rearrange slightly, interchange the order of integration on some terms, and apply the same approximations regarding the extension of the limits of integration on the variable t as in Appendix I. This gives us

$$\epsilon \approx \frac{1}{T}\int_{-\infty}^{\infty} h_e(\mu)\left\langle \frac{1}{n^2}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\int\int\int\int_{-\infty}^{\infty} h_e(\lambda)h_i(\xi, t_1 + kt_0)h_i(\nu, t_1 + lt_0)x(t-\lambda-\xi)x(t-\mu-\nu)dtd\nu d\xi d\lambda \right.$$

$$+\frac{T}{n}\frac{1}{n}\sum_{k=0}^{n-1}\int\int\int_{-\infty}^{\infty} h_e(\lambda)h_i(\xi, t_1 + kt_0)h_i(\nu, t_1 + kt_0)R_{nn}(\lambda+\xi-\mu-\nu)d\nu d\xi d\lambda$$

$$\left. -\frac{1}{n}\sum_{l=0}^{n-1}\int\int_{-\infty}^{\infty} h_i(\nu, t_1 + lt_0)h_i(t, t_1 + nt_0)x(t-\mu-\nu)dtd\nu \right\rangle_i d\mu$$

$$+\left\langle \frac{1}{T}\int_{-\infty}^{\infty} h_i^2(t, t_1 + nt_0)dt + \frac{1}{T}\int\int\int_{-\infty}^{\infty}\frac{1}{n}\sum_{l=0}^{n-1} h_e(\mu)h_i(\nu, t_1 + lt_0)h_i(t, t_1 + nt_0)x(t-\mu-\nu)dtd\nu d\mu \right\rangle_i. \quad (50)$$

Notice that the term in brackets in the first integral of (50) is identical to the bracketed term of (43) which is known to be zero. Taking ensemble averages of the separate terms of the remaining part of (50) and taking the Fourier transform of the result we thus have

$$\epsilon = \frac{1}{2\pi T}\int_{-\infty}^{\infty}\left\{\langle \mid H_i(j\omega, t_1 + nt_0)\mid^2\rangle_i - \left\langle \frac{1}{n}\sum_{k=0}^{n-1} H_i(j\omega, t_1 + kt_0)H_i^*(j\omega, t_1 + nt_0)\right\rangle_i \times H_e(j\omega)X(j\omega)\right\} d\omega. \quad (51)$$

Substituting (45) for $H_e(j\omega)$ we have the resulting expression for error

$$\epsilon = \frac{1}{2\pi T}\int_{-\infty}^{\infty}\left\{\langle \mid H_i(j\omega, t_1 + nt_0)\mid^2\rangle_i - \frac{\left|\left\langle \frac{1}{n}\sum_{k=0}^{n-1} H_i(j\omega, t_1 + kt_0)H_i^*(j\omega, t_1 + nt_0)\right\rangle_i\right|^2 \mid X(j\omega)\mid^2}{\left\langle\left|\frac{1}{n}H_i(j\omega, t_1 + kt_0)\right|^2\right\rangle_i \mid X(j\omega)\mid^2 + \frac{T}{n}\left\langle \frac{1}{n}\sum_{k=0}^{n-1}\mid H_i(j\omega, t_1 + kt_0)\mid^2\right\rangle_i S_n(j\omega)\right\rangle}\right\} d\omega. \quad (52)$$

To simplify the writing of these equations, we will use the notation

$$\overline{H_i(j\omega)} = \frac{1}{n}\sum_{k=0}^{n-1} H_i(j\omega, t_1 + kt_0).$$

Eq. (52) is the expression which must be minimized by a choice of $X(j\omega)$ subject to constraints (46) and (47)

Thus, we have

$$\delta\left[\epsilon + \lambda\left(\frac{1}{2\pi}\int_{B_1}\mid X(j\omega)\mid^2 d\omega - Wt_x\right)\right] = 0 \quad (53)$$

where λ is a parameter. Doing the operations indicated above, we have

$$0 = \frac{-1}{2\pi T}\int_{-\infty}^{\infty}\frac{\frac{T}{n}\mid\langle\overline{H_i(j\omega)}\,H_i^*(j\omega, t_1 + nt_0)\rangle_i\mid^2\langle\mid\overline{H_i(j\omega)}\mid^2\rangle_i S_n(j\omega)\delta\mid X(j\omega)\mid^2}{\left[\langle\mid\overline{H_i(j\omega)}\mid^2\rangle_i\mid X(j\omega)\mid^2 + \frac{T}{n}\langle\mid\overline{H_i(j\omega)}\mid^2\rangle_i S_n(j\omega)\right]^2} d\omega + \frac{\lambda}{2\pi}\int_{-\infty}^{\infty}\delta\mid X(j\omega)\mid^2 d\omega, \quad (54)$$

and by factoring we have

$$0 = \frac{1}{2\pi}\int_{B_1}\left\{\lambda - \frac{\frac{1}{n}\,|\,\langle\overline{H_i(j\omega)}\,H_i^*(j\omega, t_1+nt_0)\rangle_i\,|^2\langle|\,\overline{H_i(j\omega)}\,|^2\rangle_i S_n(j\omega)}{\left[\langle|\,\overline{H_i(j\omega)}\,|^2\rangle_i\,|\,X(j\omega)\,|^2 + \frac{T}{n}\,\langle|\,\overline{H_i(j\omega)}\,|^2\rangle_i S_n(j\omega)\right]^2}\right\}\delta|\,X(j\omega)\,|^2 d\omega. \tag{55}$$

The limits of integration of (55) hold by virtue of constraint—(46). Eq. (55) cannot hold unless the term in braces is zero, since $\delta|\,X(j\omega)\,|^2$ is an arbitrary function. Setting this term to zero and solving the resulting equation for $|\,X(j\omega)\,|^2$ as a function of λ we have

$$|\,X(j\omega)\,|^2 = G^{1/2}(j\omega)S_n{}^{1/2}(j\omega)\left[\left(\frac{|\,\langle\overline{H_i(j\omega)}H_i^*(j\omega, t_1+nt_0)\rangle_i\,|^2}{n\lambda\langle|\,\overline{H(_ij\omega)}\,|^2\rangle_i}\right)^{1/2} - \frac{T}{n}G^{1/2}(j\omega)S_n{}^{1/2}(j\omega)\right], \qquad \omega \text{ in } B_2$$

$$= 0, \qquad \omega \text{ in } \overline{B}_2. \tag{56}$$

$G(j\omega)$ is defined as

$$G(j\omega) = \frac{\overline{\langle|\,H_i(j\omega)\,|^2\rangle_i}}{\langle|\,\overline{H_i(j\omega)}\,|^2\rangle_i}.$$

B_2 is that sub-band of B_1 where $|\,X(j\omega)\,|^2 \geq 0$. This must be true since $|\,X(j\omega)\,|^2$ is an energy spectrum. Thus, instead of the optimum test signal we must work with the realizable part of the optimum test signal. $\overline{B}_2$ is of course the remainder of the frequency spectrum.

From (47) and (56), we obtain a solution for λ. Thus,

$$Wt_x = \frac{1}{2\pi}\int_{B_2} G^{1/2}(j\omega)S_n{}^{1/2}(j\omega)\left(\frac{|\,\langle\overline{H_i(j\omega)}\,H_i^*(j\omega, t_1+nt_0)\rangle_i\,|^2}{\overline{\langle|\,H_i(j\omega)\,|^2\rangle_i}}\right)^{1/2}\frac{1}{\sqrt{n\lambda}}\,d\omega - \frac{1}{2\pi}\int_{B_2}\frac{T}{n}\,G(j\omega)S_n(i\omega)d\omega \tag{57}$$

from which we get

$$\sqrt{\lambda} = \frac{\frac{\sqrt{n}}{2\pi}\int_{B_2} G^{1/2}(j\omega)S_n{}^{1/2}(j\omega)\,|\,\overline{H_i(i\omega)}\,H_i^*(j\omega, t_1+nt_0)\,|\,(\langle|\,\overline{H_i(j\omega)}\,|^2\rangle_i)^{-1/2}d\omega}{nWt_x + T\frac{1}{2\pi}\int_{B_2} G(j\omega)S_n(j\omega)d\omega}. \tag{58}$$

Finally, we have for the optimum test signal

$$X(j\omega) = G^{1/4}(j\omega)S_n{}^{1/4}(j\omega)\left[\left(\frac{|\,\overline{H_i(j\omega)}\,H_i^*(j\omega, t_1+nt_0)\,|^2}{n\lambda\langle|\,\overline{H_i(j\omega)}\,|^2\rangle_i}\right)^{1/2} - \frac{T}{n}G^{1/2}(j\omega)S_n{}^{1/2}(j\omega)\right]^{1/2}e^{-j\beta(\omega)}, \qquad \omega \text{ in } B_2,$$

$$= 0, \qquad \omega \text{ in } B_2, \tag{59}$$

where $\beta(\omega)$ is an arbitrary phase function.

Acknowledgment

The author would like to express appreciation to Dr. M. E. Van Valkenburg, Dr. J. B. Cruz, Jr., and the members of the Adaptive Systems Group at the Coordinated Science Laboratory, University of Illinois, Urbana.

Bibliography

[1] J. S. Bendat, "Principles and Applications of Random Noise Theory," John Wiley and Sons, Inc., New York, N. Y.; 1958.
[2] Y. W. Lee, "Applications of Statistical Methods to Communications Problems," Res. Lab. of Electronics, Mass. Inst. Tech., Cambridge, Tech. Rept. 181, pp. 25–28; September, 1950.
[3] IRE Trans. on Information Theory, vol. IT-6; June, 1960.
[4] G. L. Turin, "On the estimation in the presence of noise of the impulse response of a random, linear filter," IRE Trans. on Information Theory, vol. IT-3, pp. 5–10; March, 1957.
[5] J. L. Lawson, and G. E. Uhlenbeck, "Threshold Signals," M.I.T. Rad. Lab. Ser., vol. 24, McGraw-Hill Book Co., Inc., New York, N. Y.; 1953.
[6] W. W. Lichtenberger, Ph.D. Dissertation, University of Illinois, Urbana, Ill., 1961. Also "The Identification of Linear Processes by Means of Correlating Filters," Coordinated Science Lab., University of Illinois, Urbana, Tech. Rept. No. R-122; December, 1960.
[7] G. L. Turin, "An introduction to matched filters," IRE Trans. on Information Theory, vol. IT-6, pp. 311–329; June, 1960.
[8] E. A. Huber, "A Method of Adaptive Control for High-Order Systems," Coordinated Science Lab., University of Illinois, Urbana, Tech. Rept. No. R-121; August, 1960.

A Modified Lyapunov Method for Nonlinear Stability Analysis*

D. R. INGWERSON†, member, ire

Summary—**A modification of the original Lyapunov stability criterion is given, which includes intermediate conditions of stability, as well as "stability in the small" and "stability in the large." A means is developed for applying it to practical control problems that eliminates much of the guesswork usually required with Lyapunov methods of investigation. The process is based upon an integration of matrices which solves the linearized problem exactly. It gives sufficient conditions for the stability of nonlinear systems that are always correct for small disturbances and may be exact or conservative for large deviations from equilibrium. The formal procedure admits to enough variation that a wide range of nonlinear problems can be treated. Examples from both continuous and discontinuous feedback systems are given to illustrate its use.**

Introduction

In recent years the demands for a means of nonlinear stability analysis have caused investigators of automatic control problems to divert their attention from the various transform methods to some of the more fundamental ways of treating dynamic systems. Most of the well-known procedures of dynamics have been of little use because they were devised for a different kind of problem. One classical means of analysis, the second method of Lyapunov, is becoming increasingly recognized as having great potentiality, both for resolving nonlinear stability and performance problems and for providing a different philosophical viewpoint for the general control field.

The fundamental concepts of the method appeared in a Russian publication in 1892. As presented by Lyapunov it was not supposed to handle the kind of problems associated with automatic control, but the provision for modification and extension was evident, and to some extent indicated, in the original statement. Modern researchers, mostly in Russia, have been concerned with developing a general theory of control using this as a basis.

The Lyapunov method is essentially a stability criterion which uses a generalization of Lagrange's theorem of minimum potential energy to establish conditions for equilibrium. The principal device that provides this information is a function, called the Lyapunov function, which varies in magnitude according to the state of a system. It is the generation or induction of these functions that poses the major problem in the process.

The Lyapunov stability criterion deals only with arbitrarily small disturbances. While this is adequate for linear systems and various applications in mechanics, nonlinear control problems require a knowledge of the behavior of systems in response to large disturbances. A generalization of the original theorem which applies to arbitrarily large and arbitrarily small disturbances and to intermediate conditions as well is given in this paper.

In contrast to the success that has been achieved in advancing the theoretical concepts of stability by this

* Received by the PGAC, December 10, 1960; revised manuscript received, March 15, 1961. This paper is based upon "A Modified Lyapunov Method for Nonlinear Stability Problems," Ph.D. dissertation, Stanford University, Stanford, Calif.; November, 1960. This dissertation gives a more detailed discussion of the theory and applications.

† Sunnyvale Development Center of the Sperry Gyroscope Co., Sunnyvale, Calif.

method, little has been accomplished in the way of formulating practical means for applying them to specific problems. This is due to the lack of formal processes for producing Lyapunov functions. However, many problems that could be treated in no other way have been resolved by a fortunate combination of manipulation and guesswork.

A method which is easily applied to many of the systems encountered in automatic control and which has given good results for numerous examples is presented here. It is based on the observation that the quadratic part of the Lyapunov function must give stability information for the linearized equation of motion of a system. The procedure permits several modifications to cover a broad range of the circumstances that can occur in the general control problem.

I. Representation of Control Systems

For the concepts of stability and control of a system to have any meaning, the system must be capable of changing its state, in some sense of the word; and if it possesses this property, there is some mathematical means for describing the nature of the change. For the systems considered here, this is the differential equation of motion. Specifically, this discussion is concerned with a system of the type shown in Fig. 1. It is a plant—a dynamical system or otherwise—whose outputs are to be regulated by means of a set of inputs which are compared to certain output variables. The differences are operated upon by a controller, which in turn supplies signals to the plant. In addition, both the plant and controller may be subjected to the influence of various uncontrolled or free inputs. These represent such things as temperature variations, component aging or other parametric excitation. This is essentially a conventional control system.

It is assumed that the system can be represented by a normal set of equations of the form:

$$\dot{x}_i = f_i(x_j, x_k - r_k(t), u_s(t)) \qquad \begin{matrix} i, j = 1, 2, \cdots, n \\ k = 1, 2, \cdots, p \\ s = 1, 2, \cdots, q. \end{matrix} \quad (1)$$

While this can always be accomplished in theory it is easy to conceive of cases where the solution of high-order or transcendental equations is necessary to produce this form. The explicit dependence of the equations on time is broken into two parts, the *reference* inputs $r_k(t)$ and the *free* inputs $u_s(t)$. The x_j represent the magnitudes of n different variables which completely specify the state of the plant and controller. They are called the *state variables* or, from a geometrical viewpoint, they form the components of the *state vector*.

Any condition where all of the state variables are constant is called an *equilibrium position*. This is characterized by all of the functions f_i in (1) being equal to zero. The dependence on time of a physical system may be divided into those factors which influence an equilibrium position and those which do not. It is assumed that the $r_k(t)$ affect an equilibrium position of (1) while the $u_s(t)$ do not. Thus the $r_k(t)$ may include factors that are not specifically intended as reference inputs. The equilibrium position can exist only when the $r_k(t)$ are constant.

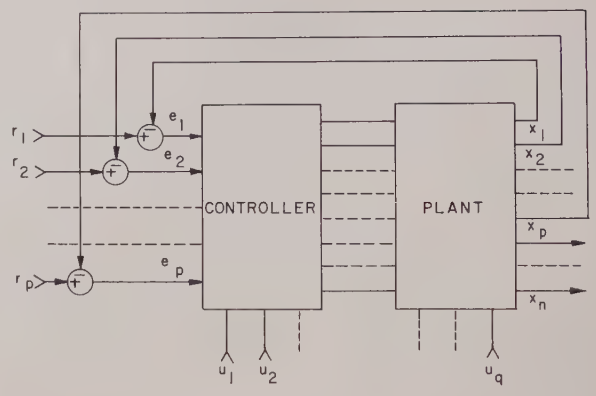


Fig. 1—General control system.

The equation describing the motion has a *solution* defined as any set of functions $\phi_j(t)$ which, if they are substituted for the state variables x_j, satisfy the equations of motion. Solutions are dependent on time in the manner prescribed by the equations of motion and also upon an initial state x_{j0} and the time t_0 when this state occurs. Thus the notation $\phi_j(t)$ is understood to imply a set of functions which have the value x_{j0} at a specified time t_0; $\phi_j(t) = \phi_j(t_0, x_{j0}; t)$.

In most practical cases the equations can be expressed in the somewhat simpler form:

$$\dot{x}_i = f_i(x_j - x_j^0(t), u_s(t)) \qquad \begin{matrix} i, j = 1, 2, \cdots, n \\ s = 1, 2, \cdots, q. \end{matrix} \quad (2)$$

$x_j^0(t)$ is defined as $r_j(t)$ for $j \leq p$ and zero for $j > p$. For convenience it is assumed that the equations are set up so that $f_i(0, u_s(t)) = 0$; that is, there are no constant terms in the functions. This does not place any restrictions on the system but only specifies the manner of writing the equations. With this mathematical description, the system of Fig. 1 can be represented by the vector block diagram of Fig. 2. Systems of this kind are the best suited for treatment by the Lyapunov method.

To interpret the concepts of the Lyapunov method, the idea of an n-dimensional state space is used. Each of the state variables is imagined to represent a length along an axis in this hyperspace. If the variables are functionally independent, no three axes lie in a plane. Under certain circumstances, the equation of motion of a system may be set up as a single differential equation of nth order.

$$\frac{d^n y}{dt^n} + g\left(\frac{d^{n-1} y}{dt^{n-1}}, \cdots, \frac{dy}{dt}, y, t\right) = 0. \quad (3)$$

With the substitutions $x_1 = y$, $x_2 = dy/dt$, etc., a special form of (1) or (2) is immediately produced. In this case

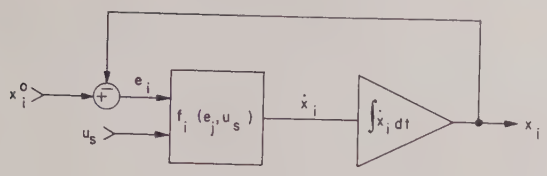


Fig. 2—Equivalent control system.

the state space reduces to the conventional phase space in which each succeeding axis represents the rate of change of the quantity measured along the one preceding it.

Any point in state space is represented by a vector having the various state variables as components. The totality of all points in the space represents all possible states which the system may assume. With the passage of time the state vector traces a curve in the space, known as a system trajectory. Fig. 3 is a possible trajectory in three-dimensional space. It may be thought of as the projection of a solution of the equations, plotted in a space containing the n-state axes and a time axis, onto the state space. Such a space and the projection is shown in Fig. 4.

II. Definitions of Stability

Solutions of equations may be divided into general classes according to whether or not they possess certain properties. The class into which a solution falls can often be determined without knowledge of the solution. One useful division is into the classes *stable* and *unstable*. The primary purpose of the Lyapunov method is to furnish a criterion by which an investigator can decide into which of these categories the solutions of a particular system fall.

To make the decision, a definition of stability is required. Various definitions of stability for a specified state of a dynamic system are available. The concept is somewhat arbitrary depending on the particular requirements for the system. A set of definitions for various applications is presented by Ingwerson [2]. Some of these are repeated here.

Stability in the Sense of Lyapunov

Let $\phi_i(t)$ be a solution of the set of equations,

$$\dot{x}_i = f_i(x_i, t), \tag{4}$$

and define a new set of variables; $q_i = x_i - \phi_i$. If these are substituted into (4), a new set of equations results which has an equilibrium position at $\boldsymbol{q} = \mathbf{0}$. This is true whether the original equations possess an equilibrium position or not.

The equilibrium position is called *stable in the sense of Lyapunov* if for every $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that $\|\boldsymbol{q}(t)\| < \epsilon$ whenever $\|\boldsymbol{q}(t_0)\| < \delta$ for all $t > t_0$. In other words, the equilibrium position, $\boldsymbol{q} = \mathbf{0}$, is stable if the magnitude of the state vector $\boldsymbol{q}$ can be made to remain below an arbitrary upper bound by choosing for it a sufficiently small initial magnitude. If $\boldsymbol{q}$ approaches zero as t approaches infinity the equilibrium position is called asymptotically stable in this sense.

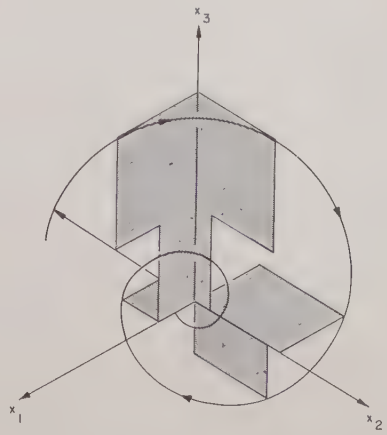


Fig. 3 -System trajectory.

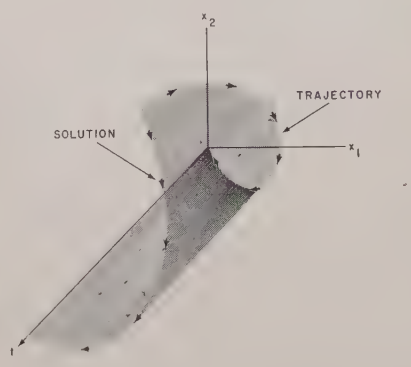


Fig. 4—Projection of a solution.

More General Definitions of Stability

The preceding definition applies only to the behavior of solutions in an arbitrarily small region of state space about the equilibrium position. It is sometimes called "stability in the small" or A stability.

Let the reference inputs to the system of (1) be constant. Then a broader definition of stability is stated: *an amplitude position of* $\mathbf{x}_e$ *of* (1) *is stable in a region* $\boldsymbol{R}$ *in state space at a time* t_0 *if for every* $\epsilon > 0$ *there exists a* $\delta(\epsilon) > 0$ *such that* $\|\boldsymbol{\phi}_1(t_0) - \boldsymbol{\phi}_2(t_0)\| < \delta$ *implies* $\|\boldsymbol{\phi}_1(t) - \boldsymbol{\phi}_2(t)\| < \epsilon$ *for all* $\boldsymbol{\phi}_1(t_0)$ *and* $\boldsymbol{\phi}_2(t_0)$ *in* $\boldsymbol{R}$ *and all* $t > t_0$.

This definition eliminates the three unstable phenomena, unbounded variations, other equilibrium states, and limit cycles from those solutions which satisfy its requirements. The equilibrium position is called *asymptotically stable* if in addition $\boldsymbol{\phi}(t) \to \mathbf{x}_e$ as $t \to \infty$. It is called *neutrally stable* if it is stable and not asymptotically stable.

The region $\boldsymbol{R}$ may be specified as arbitrarily small, finite, or arbitrarily large. If it is of a specified finite size the solutions are called B *stable* or "stable in the large." If it may be chosen arbitrarily large the solutions are variously said to be B_∞ *stable*, "stable in the large," "stable in the whole" or "globally stable." For convenience the prefixes B and B_∞ are used here.

Stability of Systems with Forcing Functions

When the $r_k(t)$ in (1) are not constant, the stability of an equilibrium position can no longer be considered. However, if the $r_k(t)$ were held constant at any time, a corresponding equilibrium position would exist. Thus a time-varying set of references may be thought of as an effective movement of an equilibrium state from point to point in state space.

Let $\mathbf{x}_e(t)$ be the equilibrium position corresponding to a set of reference inputs $r_k(t)$ and let $\boldsymbol{\phi}(t_0)=\mathbf{x}_e(t_0)$. A system is *stable at t_0 with respect to this set of inputs* if $\boldsymbol{\phi}(t_1)$ falls inside of a region for which the equilibrium position corresponding to $r_k(t_1)$ is stable by the previous definitions for all $t_1>t_0$. Evidently, if a system is to be stable for all possible inputs, all equilibrium positions $\mathbf{x}_e(t_1)$ must be B_∞ stable.

III. Stability Criteria

The preceding section gives several definitions of stability. Now it is desired to find criteria and general methods for applying them whereby the stability of specific systems can be determined from a consideration of the differential equations of motion. The definitions supply both necessary and sufficient conditions for a system to have stable solutions; however, a criterion for stability, based on a definition, may provide only a sufficient condition. This is the case in most applications of the Lyapunov method.

The Second Method of Lyapunov

The basis for the procedures outlined here, as well as for a variety of other modifications, is the original second, or direct, method of Lyapunov. It was intended to determine the stability of a position or state of motion of a dynamic system in response to small disturbances and, therefore, is valid only for "stability in the sense of Lyapunov." A complete statement of the stability theorem is found in Lyapunov [3] and is repeated in more modern terminology by Hahn [4], Kalman and Bertram [5], and Cunningham [6].

The criterion applies directly to the equations of the disturbed motion derived from (4) by the substitution, $q_i=x_i-\phi_i(t)$. The new equations are:

$$\dot{q}_i = f_i(q_j+\phi_j(t),\, t) - f_i(\phi_j(t),\, t) = \theta_i(q_j,\, t). \tag{5}$$

The Lyapunov criterion for the stability of (5) can be stated as follows: *If there exists a positive definite function $V(q_i, t)$, continuous in q_i, such that its derivative with respect to time, $\dot{V}(q_i, t)$, is negative or zero for all $t>t_0$, the system is stable in the sense of Lyapunov.*

V is called a Lyapunov function for the system. $V(q_i, t)$ is defined to be positive definite if it is positive for all $\|\mathbf{q}\|>0$ and all $t>t_0$ and is zero for $\mathbf{q}=\mathbf{0}$. For example, $q_1^2+q_2^2\mathrm{Log}\ t$, is positive definite for $t>1$ while, $q_1^2+q_2^2e^{-t}$, is not positive definite since the coefficient of q_2^2 becomes less than any positive constant. Functions which are always positive or zero are called "non-negative" or "positive semi-definite."

The derivative of V is to be taken along a trajectory of the motion. That is:

$$\dot{V} = \frac{\partial V}{\partial q_1}\theta_1(q_j,\, t) + \frac{\partial V}{\partial q_2}\theta_2(q_j,\, t) + \cdots + \frac{\partial V}{\partial q_n}\theta_n(q_j,\, t) + \frac{\partial V}{\partial t}. \tag{6}$$

$\dot{V}$ is then a function of the components of $\boldsymbol{q}$, although not necessarily all of them, and time; it must be either negative or zero at every point in state space for stability. Clearly, the result would be the same if V were negative definite and $\dot{V}$ were positive or zero, and so the terms "negative" and "positive" may be interchanged in the statement.

This is a valid criterion for stability but only for arbitrarily small deviations from equilibrium. To show that it is not sufficient to determine stability for large disturbances an example of Barbasin and Krasovskii [7] is cited. Let a system be described by the following equations:

$$\dot{x}_1 = -\frac{6x_1}{(1+x_1^2)^2} + 2x_2, \qquad \dot{x}_2 = -\frac{2(x_1+x_2)}{(1+x_1^2)^2}. \tag{7}$$

That the system is stable for small disturbances may be seen by linearization about the point $\mathbf{x}=\mathbf{0}$; however, the slope of the trajectories along the hyperbola

$$x_2 = \frac{2}{x_1-\sqrt{2}}$$

is,

$$(dx_2/dx_1)_T = \frac{-1}{1+2\sqrt{2}\,x_1+2x_1^2}.$$

This is greater than the slope of the hyperbola,

$$(dx_2/dx_1)_C = -\frac{1}{1-\sqrt{2}\,x_1+\dfrac{x_1^2}{2}}, \quad \text{for} \quad x_1>\sqrt{2}.$$

Since $\dot{x}_1$ is positive along this curve, all system trajectories are going into the region bounded by it and none are leaving. In any region $\boldsymbol{R}$ that includes a part of this region, the system is unstable.

But assume as a Lyapunov function the positive definite function given by the following relation:

$$V = \frac{x_1^2}{1+x_1^2} + x_2^2.$$

The derivative along a trajectory of (7) is:

$$\dot{V} = -\frac{12x_1^2}{(1+x_1^2)^4} - \frac{4x_2^2}{(1+x_1^2)^2},$$

which is negative definite. These conditions imply stability by the above criterion but only for arbitrarily small disturbances.

A Modified Lyapunov Criterion

To investigate the stability of systems represented by (2) for finite disturbances, let the references $x_j^0(t)$ be fixed at a constant value at some time t_1 in accordance with the definition given for the stability of systems with forcing functions. Define, $q_i = x_i - x_i^0(t_1)$, for $t > t_1$. Eq. (2) then becomes

$$\dot{q}_i = f_i(q_j, u_s(t)). \tag{8}$$

The following theorem gives a stability criterion for (8): *The equilibrium position, $\mathbf{q} = \mathbf{0}$, is stable in a region $\boldsymbol{R}$ at a time t_0 if there exists a continuous function, $V(q_i, t)$, which is zero at the equilibrium position and is greater than some positive definite function, $W(q_i)$, for $t > t_0$, elsewhere in the region, and such that:* 1) *one of the surfaces, $V =$ a constant, bounds the region,* 2) *the gradient of the function, ∇V, is not zero anywhere in the region except at the equilibrium position and* 3) *the derivative of the function with respect to time, $\dot{V}(q_i, t)$, is negative or zero inside of the region.*

The quantity ∇V is a vector with components $\partial V/\partial q_i$, all of which cannot be zero simultaneously in $\boldsymbol{R}$ except at the equilibrium position. This is equivalent to a statement that the function cannot have a relative maximum inside the region. In general, when the $u_s t$ are not constant, the boundary of the stable region varies with time. The above conditions assure that if $\boldsymbol{q}$ falls inside of $\boldsymbol{R}$ at any time it will remain permanently inside of some bounded region.

To prove this theorem observe that in the neighborhood of $\boldsymbol{q} = \mathbf{0}$ there is some surface, $V(q_i, t) =$ a positive constant, which completely encloses the point $\boldsymbol{q} = \mathbf{0}$. This follows from the requirement that V be greater than a positive definite function. At the same time another surface for which V is equal to a larger positive constant exists which encloses the former one. Thus a series of hypersurfaces exist in state space, characterized by a series of constants which increase monotonically as the distance from the origin increases. These continue out until a point is reached where the gradient of the function vanishes. Surfaces beyond this point may correspond to smaller constant values than for the surface passing through the point.

The requirement that $\dot{V}$ be nonpositive assures that the system trajectories in $\boldsymbol{R}$ either remain on surfaces of constant V or cross them toward the equilibrium position. In time-varying systems the position of a surface $V(q_i, t) = V_0$ may change but it is always bounded by a surface $W(q_i) = W_0$ since $V > W$. Similarly a surface $V(q_i, t) = V_1$, where $V_1 < V_0$, is bounded by a surface $W(q_i) = W_1$, where W_1 can be chosen smaller than W_0. Therefore, all trajectories initially in $\boldsymbol{R}$ remain bounded and the bounds are dependent upon the initial states in accordance with the definition of stability given previously.

If this theorem is applied to the Lyapunov function given for (7), a boundary for the stable region is determined. The gradient of the function is

$$\nabla V = \left(\frac{2x_1}{(1 + x_1^2)^2}, 2x_2 \right),$$

which vanishes at the origin and at the points, $\mathbf{x} = (\pm \infty, 0)$. The equation of the boundary, $V =$ a constant, passing through these points is $x_2^2(1 + x_1^2) = 1$. This curve, shown in Fig. 5, includes an infinite region, but not all of the $x_1 x_2$ plane.

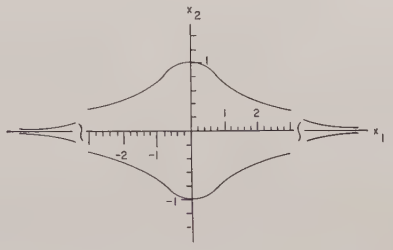


Fig. 5—A stability boundary for (7).

Since $\dot{V}$ is negative everywhere, all trajectories inside the bounded region tend toward the origin, and the solutions are asymptotically stable. Outside of the region, the solutions may or may not tend toward equilibrium. As with the original criterion, this theorem gives only sufficient conditions for stability. Another Lyapunov function may give another region of stability which includes more of the $x_1 x_2$ plane.

IV. General Results

Few general results, derived by the Lyapunov method, are yet available. Because of the specialized nature of the equations to which they apply, those which are known are more valuable for the theoretical insight they give to the method than for application to specific problems. Three of the most common results are given here to illustrate the use of the second method and to preface a discussion of more practical applications.

A Result of Lyapunov

Lyapunov gave the following result to demonstrate the second method. If the matrix of elements $\partial f_i/\partial x_j$, formed for the differential equations, $\dot{x}_i = f_i(x_j)$, is symmetric, a Lyapunov function is:

$$V(\mathbf{x}) = -\int_0^{\mathbf{x}} \boldsymbol{f}(\mathbf{x}) \cdot d\mathbf{x}.$$

This is a line integral which is uniquely integrable because the symmetry of the matrix above is the condition under which the curl of the vector $\boldsymbol{f}$ vanishes. The derivative of the function is just $-\boldsymbol{f}'\boldsymbol{f}$ which is nonpositive.[1] The condition for stability is the definite nature of V.

[1] A prime following the symbol for a vector or matrix denotes the transpose of that quantity.

B_∞ stability exists if V increases monotonically with increasing $\mathbf{x}$. A condition for this is that the eigenvalues[2] of the matrix with elements $\partial^2 V/\partial x_i \partial x_j$, which is also the matrix with elements $\partial f_i/\partial x_j$, be positive. The signs of the eigenvalues can be investigated by Sylvester's inequalities; that is, they are positive if all of the principal minors and the determinant of the matrix are positive.

A Result of Krasovskii

The result above is a special case of the following more general theorem given by Krasovskii [8]. Let $\boldsymbol{B}(\mathbf{x}, t)$ be the matrix with elements $\partial f_i/\partial x_j$ derived from the set of equations, $\dot{x}_i = f_i(x_j, t)$, where $f_i(\mathbf{0}, t) = 0$. Then B_∞ stability exists if the eigenvalues of the symmetric part of $\boldsymbol{B}$ (*i.e.*, of $\boldsymbol{B}' + \boldsymbol{B}$) are negative. This is proved (see [4] and [5]) for the autonomous case with a Lyapunov function, $V = \mathbf{f}'\mathbf{f}$. It follows for the general case from an application of the autonomous Lyapunov function, $V = \mathbf{x}'\mathbf{x}$. This function obviously is monotonically increasing in $\mathbf{x}$. Its derivative is: $\dot{V} = \mathbf{x}'\mathbf{f} + \mathbf{f}'\mathbf{x}$.

To show that $\dot{V}$ is negative definite, under the hypothesis that $\boldsymbol{B}' + \boldsymbol{B}$ has negative eigenvalues, the scalar product of its gradient with the radius vector $\mathbf{x}$ is formed.

$$\nabla V \cdot \mathbf{x} = \mathbf{x}'\mathbf{f} + \mathbf{f}'\mathbf{x} + \mathbf{x}'(\boldsymbol{B}' + \boldsymbol{B})\mathbf{x} = V + \mathbf{x}'(\boldsymbol{B}' + \boldsymbol{B})\mathbf{x}.$$

It is evident that the last term on the right is a negative definite form for the hypothesized condition. Rearranging terms and dividing by the magnitude of the state vector gives

$$\frac{\nabla \dot{V} \cdot \mathbf{x} - \dot{V}}{(\mathbf{x}'\mathbf{x})^{1/2}} = \frac{\mathbf{x}'(\boldsymbol{B}' + \boldsymbol{B})\mathbf{x}}{(\mathbf{x}'\mathbf{x})^{1/2}}.$$

The left-hand side of this expression may be written as

$$\nabla\left(\frac{\dot{V}}{(\mathbf{x}'\mathbf{x})^{1/2}}\right) \cdot \mathbf{x}.$$

This quantity is negative everywhere except at the origin because it is equal to a negative definite quantity. Since it is the scalar product of the gradient of a function and a radius vector, the function $\dot{V}/(\mathbf{x}'\mathbf{x})^{1/2}$ decreases in any radial direction away from the origin. $\dot{V}$ is a form of higher order than $(\mathbf{x}'\mathbf{x})^{1/2}$, and hence this function is zero at the origin. Therefore $\dot{V}$ is negative definite.

Lyapunov's Result for Linear Autonomous Equations

The most instructive application of the second method gives necessary and sufficient conditions for the stability of linear, autonomous differential equations. The Lyapunov function, $V = \mathbf{x}'\boldsymbol{A}\mathbf{x}$, has the derivative, $\dot{V} = \mathbf{x}'(\boldsymbol{B}'\boldsymbol{A} + \boldsymbol{A}\boldsymbol{B})\mathbf{x}$, for the set of n equations

$$\dot{\mathbf{x}} = \boldsymbol{B}\mathbf{x}. \tag{9}$$

[2] The eigenvalues of a matrix $\boldsymbol{M}$ are defined as the roots, λ, of the characteristic equation, Det $(\boldsymbol{M} - \lambda \boldsymbol{I}) = 0$.

$\boldsymbol{A}$ and $\boldsymbol{B}$ are $n \times n$ matrices with constant elements. For any symmetric matrix $\boldsymbol{C}$ with positive eigenvalues the equilibrium position, $\mathbf{x} = \mathbf{0}$, of the set (9) is asymptotically stable if and only if the matrix $\boldsymbol{A}$ which solves

$$\boldsymbol{B}'\boldsymbol{A} + \boldsymbol{A}\boldsymbol{B} = -\boldsymbol{C} \tag{10}$$

has positive eigenvalues. Referring to the usual stability requirement for linear, autonomous systems it is seen that $\boldsymbol{A}$ has positive eigenvalues if and only if the real parts of the eigenvalues of $\boldsymbol{B}$ are negative. Proofs of this theorem are somewhat complicated. Two different ones are found in [1] and [5].

This condition may be relaxed by permitting the derivative of the Lyapunov function to be negative semidefinite rather than negative definite. Then only one of the eigenvalues of $\boldsymbol{C}$ need be different from zero. A necessary and sufficient condition for the eigenvalues of $\boldsymbol{A}$ to be positive is still that those of $\boldsymbol{B}$ have negative real parts. For example, if the matrix element c_{22} is taken equal to unity and all other c_{ij} are zero, the solution of (10) for the second-order set

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -a_2 & -a_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

is

$$\boldsymbol{A} = \frac{1}{2a_1a_2} \begin{pmatrix} a_2 & 0 \\ 0 & 1 \end{pmatrix}.$$

A Lyapunov function and its derivative are

$$V = \frac{1}{2a_1a_2}(a_2x_1^2 + x_2^2)$$

$$\dot{V} = -\frac{x_2^2}{a_2}.$$

The vanishing of a_1, which is the condition for neutral stability, causes the function to become infinite; however, multiplication by the factor in the denominator gives a function which is valid for that case as well.

When $\boldsymbol{B}$ is derived from an equation of the form

$$y^{(n)} + a_1y^{(n-1)} + \cdots + a_{n-1}\dot{y} + a_ny = 0,$$

by the substitutions, $y = x_1$, $\dot{y} = x_2$, etc., it has the special form

$$\boldsymbol{B} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{pmatrix}. \tag{11}$$

A set of the matrices $\boldsymbol{A}$ for systems up to the fourth order are shown in Table I. These are derived by setting successive elements on the principal diagonal of $\boldsymbol{C}$ equal to unity while the remaining elements are zero and solving (10) when $\boldsymbol{B}$ is of the form given above. In each case a factor appears in the denominator. Both of the

TABLE I

MATRICES OF LYAPUNOV FUNCTIONS

Second Order

$$A_1 = \begin{pmatrix} a_2 & 0 \\ 0 & 1 \end{pmatrix} \qquad C_1 = \begin{pmatrix} 0 & 0 \\ 0 & 2a_1 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} a_1^2 + a_2 & a_1 \\ a_1 & 1 \end{pmatrix} \qquad C_2 = \begin{pmatrix} 2a_1a_2 & 0 \\ 0 & 0 \end{pmatrix}$$

Third Order

$$A_1 = \begin{pmatrix} a_3^2 & a_2a_3 & 0 \\ a_2a_3 & a_1a_3 + a_2^2 & a_3 \\ 0 & a_3 & a_2 \end{pmatrix} \qquad C_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2(a_1a_2 - a_3) \end{pmatrix}$$

$$A_2 = \begin{pmatrix} a_1a_3 & a_3 & 0 \\ a_3 & a_1^2 + a_2 & a_1 \\ 0 & a_1 & 1 \end{pmatrix} \qquad C_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2(a_1a_2 - a_3) & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} a_1a_2^2 - a_2a_3 + a_1^2a_3 & a_1^2a_2 & a_1a_2 - a_3 \\ a_1^2a_2 & a_1^3 + a_3 & a_1^2 \\ a_1a_2 - a_3 & a_1^2 & a_1 \end{pmatrix} \qquad C_3 = \begin{pmatrix} 2a_3(a_1a_2 - a_3) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Fourth Order

$$R_2 = a_1a_2 - a_3 \qquad R_3 = a_1a_2a_3 - a_3^2 - a_1^2a_4$$

$$A_1 = \begin{pmatrix} a_3a_4^2 & a_4a_3^2 & a_2a_3a_4 - a_1a_4^2 & 0 \\ a_4a_3^2 & a_1a_4^2 + a_3^3 & a_2a_3^2 & a_3a_4 \\ a_2a_3a_4 - a_1a_4^2 & a_2a_3^2 & a_1a_3^2 + a_2^2a_3 - a_1a_2a_4 - a_3a & a_3^2 \\ 0 & a_3a_4 & a_3^2 & a_2a_3 - a_1a_4 \end{pmatrix}$$

$$C_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2R_3 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} a_1a_4^2 & a_1a_3a_4 & a_3a_4 & 0 \\ a_1a_3a_4 & a_1a_3^2 + a_4R_2 & a_3^2 + a_1^2a_4 & a_1a_4 \\ a_3a_4 & a_3^2 + a_1^2a_4 & a_1^2a_3 + a_2a_3 - a_1a_4 & a_1a_3 \\ 0 & a_1a_4 & a_1a_3 & a_3 \end{pmatrix}$$

$$C_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2R_3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} a_4R_2 & a_1^2a_4 & a_1a_4 & 0 \\ a_1^2a_4 & a_2R_2 - a_1a_4 + a_1^2a_3 & a_1^2a_2 & R_2 \\ a_1a_4 & a_1^2a_2 & a_1^3 + a_3 & a_1^2 \\ 0 & R_2 & a_1^2 & a_1 \end{pmatrix}$$

$$C_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2R_3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_4 = \begin{pmatrix} a_3R_3 + a_4(a_2R_2 - a_1a_4) & a_3(a_2R_2 - a_1a_4) & a_1R_3 + a_4R_2 & R_3 \\ a_3(a_2R_2 - a_1a_4) & (a_2^2 - a_4)R_2 + a_1a_4(a_1^2 - a_2) & a_1a_2R_2 & a_2R_2 - a_1a_4 \\ a_1R_3 + a_4R_2 & a_1a_2R_2 & a_1^2R_2 + a_1a_4 & a_1R_2 \\ R_3 & a_2R_2 - a_1a & a_1R_2 & R_2 \end{pmatrix}$$

$$C_4 = \begin{pmatrix} 2a_4R_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ in the table are multiplied by this factor.

When $\boldsymbol{B}$ appears in the general form the same process can be used to find Lyapunov functions, but the task of computation is more tedious. Ingwerson [1] gives a simpler method for deriving these. The two matrices of second order for the general case are

$$\boldsymbol{A}_1 = \begin{pmatrix} b_{21}^2 & -b_{11}b_{21} \\ -b_{11}b_{21} & b_{11}^2 + b_{11}b_{22} - b_{12}b_{21} \end{pmatrix}$$

$$\boldsymbol{A}_2 = \begin{pmatrix} b_{22}^2 + b_{11}b_{22} - b_{12}b_{21} & -b_{12}b_{22} \\ -b_{12}b_{22} & b_{12}^2 \end{pmatrix}.$$

V. Generation of Lyapunov Functions

Lyapunov's result for linear, autonomous equations gives a means for checking any general method for deriving Lyapunov functions. A general process should give the same results as are achieved by that method when it is applied to linear problems. There are, of course, any number of ways this could be accomplished, most of which would not give necessary and sufficient conditions for nonlinear systems. The Lyapunov method, however, is useful largely because it requires only sufficient conditions for stability. It is impossible to find the exact solution of most nonlinear problems by analytic means. Thus the conditions derived are nearly always conservative estimates.

It is considerably less difficult to make a good approximation for a Lyapunov function for a particular equation than to approximate a solution because, except for neutrally stable systems, Lyapunov functions are not unique. As is apparent from the previous discussion a great many functions which solve linear problems can be found by varying the matrix $\boldsymbol{C}$.

If the set of autonomous equations

$$\dot{x} = \boldsymbol{f}(\mathbf{x}) \tag{12}$$

are differentiated there results

$$\ddot{\mathbf{x}} = \boldsymbol{B}(\mathbf{x})\dot{\mathbf{x}}$$

where $\boldsymbol{B}(\mathbf{x})$ is the matrix with elements $\partial f_i/\partial x_j$. Eq. (10) may be solved for a matrix $\boldsymbol{A}(\mathbf{x})$ if the constant matrix, $\boldsymbol{B}$, is replaced by the variable one, $\boldsymbol{B}(\mathbf{x})$. In the linear case it may be observed that $\boldsymbol{A}$ is the matrix with elements $\frac{1}{2}(\partial^2 V/\partial x_i \partial x_j)$. If this were also true for the matrix $\boldsymbol{A}(\mathbf{x})$, a Lyapunov function could be found by performing two integrations. The procedure would give the correct answers to linear problems as is desired of a general method.

Certain conditions are necessary, however, for the elements of a matrix to be the second partial derivatives of a scalar function. First, the matrix must be symmetric, which is true of $\boldsymbol{A}(\mathbf{x})$, but also the relation

$$\frac{\partial a_{ij}}{\partial x_k} = \frac{\partial a_{ik}}{\partial x_j}$$

must exist for the elements of $\boldsymbol{A}(\mathbf{x})$. The latter condition is not true in general for matrices so derived. Nevertheless, $\boldsymbol{A}(\mathbf{x})$ has a desirable property; the Lyapunov function derived from $\boldsymbol{A}(\mathbf{0})$ is valid in the vicinity of the origin. It gives correct results for the linearized part of (12) and is, therefore, a first approximation. If the quadratic terms in its derivative do not predict stability, the system is unstable.

A simple means of forming a Lyapunov function from $\boldsymbol{A}(\mathbf{x})$ often gives good results. A matrix $\boldsymbol{A}(x_i, x_j)$ is formed by letting all of the variables in each element, a_{ij}, of $\boldsymbol{A}(\mathbf{x})$ vanish except x_i and x_j, where i and j are the respective indexes of the row and column containing the element. The gradient of a scalar function is then found by performing the following integration:

$$\nabla V = \int_0^{\mathbf{x}} \boldsymbol{A}(x_i, x_j)d\mathbf{x}. \tag{13}$$

This is not a line integral; the integration is to be performed over each component of $\mathbf{x}$ as if the remaining variables were constant. The result is a vector with n components whose curl is zero and, hence, is the gradient of some scalar function. A more precise definition of (13) and a proof that the vector is a gradient can be given with the aid of tensor notation. These are found in Ingwerson [1]. An example will serve to illustrate the meaning here.

The matrix $\boldsymbol{B}(\mathbf{x})$ for the third-order system described by

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = x_3$$

$$\dot{x}_3 = -(x_1 + cx_2)^3 - bx_3$$

is derived as discussed above.

$$\boldsymbol{B}(\mathbf{x}) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -3(x_1 + cx_2)^2 & -3c(x_1 + cx_2)^2 & -b \end{pmatrix}.$$

This has the same form as (11), but with variable elements. The third-order matrix corresponding to $\boldsymbol{A}_2$ in Table I is

$$\boldsymbol{A}_2(\mathbf{x}) = \begin{pmatrix} 3b(x_1 + cx_2)^2 & 3(x_1 + cx_2)^2 & 0 \\ 3(x_1 + cx_2)^2 & b^2 + 3c(x_1 + cx_2)^2 & b \\ 0 & b & 1 \end{pmatrix}.$$

It is the same as $\boldsymbol{A}_2$ except that the constant elements are replaced by corresponding variable ones.

The matrix $\boldsymbol{A}_2(x_i, x_j)$ is produced by letting certain of the variables in $\boldsymbol{A}_2(\mathbf{x})$ vanish.

$$\boldsymbol{A}_2(x_i, x_j) = \begin{pmatrix} 3bx_1^2 & 3(x_1 + cx_2)^2 & 0 \\ 3(x_1 + cx_2)^2 & b^2 + 3c^3x_2^2 & b \\ 0 & b & 1 \end{pmatrix}.$$

Performing the integration indicated by (13) yields the following gradient:

$$\nabla V = \begin{bmatrix} bx_1^3 + \frac{1}{c}(x_1 + cx_2)^3 - \frac{x_1^3}{c} \\ (x_1 + cx_2)^3 + b^2x_2 + bx_3 \\ bx_2 + x_3 \end{bmatrix}.$$

A Lyapunov function is found by evaluating the following line integral along any path of integration:

$$V = \int_0^x \nabla V \cdot d\mathbf{x} = \frac{bx_1^4}{4} + \frac{(x_1 + cx_2)^4}{4c} - \frac{x_1^4}{4c} + \frac{b^2x_2^2}{2} + bx_2x_3 + \frac{x_3^2}{2}.$$

When b, c and the quantity $bc-1$ are positive, V is positive definite and its gradient vanishes only at $\mathbf{x}=\mathbf{0}$. This can be verified by examining the zeros of the gradient, by applying Sylvester's inequalities to the matrix whose elements are $\partial^2 V/\partial x_i \partial x_j$, or just by inspection of the function.

The derivative function is determined by forming the scalar product of the gradient and the derivative of the state vector:

$$\dot{V} = \nabla V \cdot \dot{\mathbf{x}} = -(bc-1)(3x_1^2 + 3cx_1x_2 + c^2x_2^2)x_2^2.$$

It is negative semidefinite under the same conditions for which the Lyapunov function is positive. Stability of the equations of motion is therefore assured for $b>0$, $c>0$, and $bc-1>0$. In this case, the conditions are also necessary, since if $bc-1$ is negative, there is still some part of state space arbitrarily near to the origin in which V is positive. $\dot{V}$ is also positive there and hence some trajectories must diverge from equilibrium.

The above procedure thus produces a Lyapunov function which yields both necessary and sufficient conditions for this particular example. It is not, however, a strictly formal method that may be applied to any problem with assurance of valid results. The selection of the matrix $\boldsymbol{A}(x)$ requires a certain amount of ingenuity. If either of the other two matrices in Table I were used, the same result would not be achieved. The matrices in Table I are each designed to give one quadratic term in the derivative function when the system is linear. Often only one of these gives useful information for a nonlinear system, as is true in the preceding example. More frequently, it is necessary to form some linear combination of Lyapunov functions derived from the various matrices. Occasionally, it requires a matrix which is determined by selecting some of the off-diagonal elements of $\boldsymbol{C}$ in (10) to be other than zero. These are not given in Table I but are easily calculated.

When $\boldsymbol{A}(x)$ is determined, the matrix $\boldsymbol{A}(x_i, x_j)$ is defined by neglecting certain variables in each element. It is often apparent after a function has been generated that a modification of the matrix would produce more desirable results. Only one kind of modification is considered. This amounts to dropping more of the variable terms in the matrix, thereby simplifying the resulting functions. The integrability conditions are not altered by this variation provided that the symmetry of the matrix is retained.

The particular modification that has proved the most useful is applied when powers of the elements of $\boldsymbol{B}(x)$ occur in $\boldsymbol{A}(x)$. For example, the matrix $\boldsymbol{A}_1$ in Table I contains an element, a_3^2, in its first row and column. In general the corresponding $\boldsymbol{A}(x_i, x_j)$ would contain an element $a_3^2(x_1)$. It is often more desirable to replace this by the term $a_3(0)a_3(x_1)$, in which the variable is neglected in one of the factors and retained in the other.

VI. Examples from Nonlinear Feedback Systems

The method discussed in the preceding section restricts the field of possibilities from which to choose a Lyapunov function to those which can be produced from a certain class of matrices. Its most important function is to eliminate the linear aspects of a problem from consideration by giving the known correct answers for those parts. The selection of a matrix and perhaps a modification of its elements remain at the discretion of the investigator. To illustrate these points, various applications to nonlinear control are presented.

A Nonlinear Compensation

Fig. 6 shows an undamped second-order system which is made asymptotically stable by regulating the gain of an amplifier rather than by conventional linear compensation. The gain is made to vary as

$$K(e, \dot{\theta}) = b_0 - b_1 e\dot{\theta}.$$

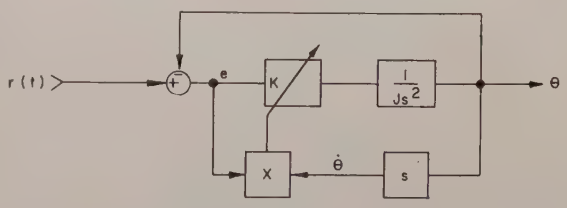


Fig. 6—A nonlinear compensation.

The equation of motion is,

$$K(e, \dot{\theta})(r - \theta) = J\ddot{\theta},$$

or with the substitutions, $\theta = x_1$, $\dot{\theta} = x_2$, $r(t) = x_1^0(t)$, and $x_2^0(t) = 0$, it is expressed in the form of (2).

$$\dot{x}_1 = x_2 - x_2^0$$

$$\dot{x}_2 = \frac{b_0}{J}(x_1 - x_1^0) - \frac{b_1}{J}(x_1 - x_1^0)^2(x_2 - x_2^0).$$

Using the definition of stability for systems with reference inputs given in the second section, x_1^0 and x_2^0 are held constant after some time t_1. Then for $t>t_1$, the

additional substitution, $y_1=x_1-x_1^0$ and $y_2=x_2-x_2^0$, reduces the above equations to

$$\dot{y}_1 = y_2$$

$$\dot{y}_2 = -\frac{b_0}{J} y_1 - \frac{b_1}{J} y_1^2 y_2.$$

If these are B_∞ stable, the system is stable for all inputs.

The matrix $\boldsymbol{B}(\boldsymbol{y})$ is

$$\begin{bmatrix} 0 & 1 \\ -\frac{b_0}{J} + \frac{2b_1}{J} y_1 y_2 & -\frac{b_1}{J} y_1^2 \end{bmatrix}.$$

The first matrix of Table I is used to form the Lyapunov function

$$\boldsymbol{A}(y_i, y_j) = \begin{bmatrix} \frac{b_0}{J} & 0 \\ 0 & 1 \end{bmatrix}.$$

From this, the procedures of the previous section give

$$V = \frac{b_0}{2J} y_1^2 + \frac{1}{2} y_2^2$$

$$\nabla V = \left(\frac{b_0}{J} y_1, y_2 \right)$$

$$\dot{V} = -\frac{b_1}{J} y_1^2 y_2^2.$$

For positive b_0 and b_1, V is positive definite and $\dot{V}$ is negative semidefinite. Unless a trajectory follows the y_1 or y_2 axis, which is obviously not possible, the system is asymptotically B_∞ stable and the compensation is accomplished.

A Discontinuous System

Fig. 7 shows a linear system whose actuating signal is switched positive or negative according to the sign of the linear switching criterion $\theta+c\dot{\theta}$. With the substitutions $\theta=x_1$ and $\dot{\theta}=x_2$, the equations of motion are

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -K\omega^2 \operatorname{sgn}(x_1 + cx_2) - \omega^2 x_1 - 2D\omega x_2.$$

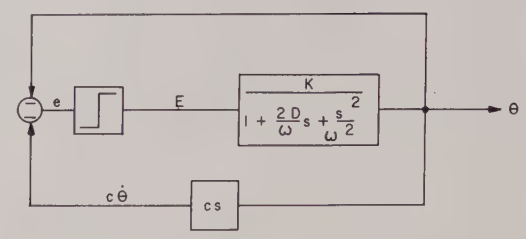


Fig. 7—Linear second-order switching.

The Dirac delta function $\delta(x)$ is defined as having infinite magnitude at $x=0$ and zero magnitude for all other values, but in such a way that

$$\int_{-\infty}^{\infty} \delta(x)dx = 1.$$

In terms of this function, it may be said that $d/dx \operatorname{sgn} x = 2\delta(x)$. Then for the equations above,

$$\boldsymbol{B}(\boldsymbol{x}) = \begin{bmatrix} 0 & 1 \\ -\omega^2 - 2K\omega^2\delta(x_1+cx_2) & -2D\omega - 2Kc\omega^2\delta(x_1+cx_2) \end{bmatrix}.$$

A combination of the two second-order matrices from Table I, $\mu\boldsymbol{A}_1+\boldsymbol{A}_2$, can be used to find a Lyapunov function. It is obvious that the element $a_{11}(\boldsymbol{x})$ must be modified, for otherwise it would contain the square of a delta function which, when integrated, leads to an infinite magnitude for the gradient. A modification in which the delta function is neglected in one of the factors of $a_1^2(\boldsymbol{x})$ and retained in the other is used:

$$\boldsymbol{A}(x_i, x_j) = \begin{bmatrix} (4D^2 + 1 + \mu)\omega^2 + 2K\omega^2(2cD\omega + 1 + \mu)\delta(x_1) & 2D\omega + 2Kc\omega^2\delta(x_1 + cx_2) \\ 2D\omega + 2Kc\omega^2\delta(x_1 + cx_2) & 1 + \mu \end{bmatrix}.$$

Integration of this matrix according to (13) gives the gradient of a function:

$$\nabla V = \begin{bmatrix} (4D^2 + 1 + \mu)\omega^2 x_1 + K\omega^2(2cD\omega + \mu) \operatorname{sgn} x_1 + 2D\omega x_2 + K\omega^2 \operatorname{sgn}(x_1 + cx_2) \\ 2D\omega x_1 + Kc\omega^2 \operatorname{sgn}(x_1 + cx_2) - Kc\omega^2 \operatorname{sgn} cx_2 + (1 + \mu)x_2 \end{bmatrix}.$$

The derivative of the function is found by combining the gradient with the original set of equations.

$$\begin{aligned} \dot{V} = &-2D\omega(\omega^2 x_1^2 + \mu x_2^2) \\ &- K^2c\omega^4[1 + \operatorname{sgn}(cx_2) \operatorname{sgn}(x_1 + cx_2)] \\ &- K\omega^2\left[(2D\omega + c\omega^2)x_1 + \left(2D\omega + \frac{\mu}{c}\right)cx_2\right] \\ &\cdot \operatorname{sgn}(x_1 + cx_2) + K\omega^2(\mu + 2cD\omega)x_2 \operatorname{sgn} x_1 \\ &+ Kc\omega^3(\omega x_1 + 2Dx_2) \operatorname{sgn} cx_2. \end{aligned}$$

While these results may give an accurate solution to the problem, they are rather complicated. An inspection of the derivative and gradient offers a means of simplification. First, if μ is set equal to ω^2c^2, one of the

terms becomes nonpositive. The terms containing sgn x_1 and sgn cx_2 complicate the derivative. From the gradient, it may be seen that these could be dropped without destroying the integrability condition. If these variations are made the Lyapunov function, its gradient and derivative are

$$V = (4D^2 + 1 + \omega^2c^2)\frac{{x_1}^2}{2} + 2D\omega x_1x_2 + K\omega^2(x_1 + cx_2)\operatorname{sgn}(x_1 + cx_2) + (1 + \omega^2c^2)\frac{{x_2}^2}{2}$$

$$\nabla V = \begin{pmatrix} (4D^2 + 1 + \omega^2c)\omega^2x_1 + 2D\omega x_2 + K\omega^2 \operatorname{sgn}(x_1 + cx_2) \\ 2D\omega x_1 + Kc\omega^2 \operatorname{sgn}(x_1 + cx_2) + (1 + \omega^2c^2)x_2 \end{pmatrix}$$

$$\dot{V} = -2D\omega^3({x_1}^2 + c^2{x_2}^2) - K^2c\omega^4 - K\omega^3(2D + c\omega)(x_1 + cx_2)\operatorname{sgn}(x_1 + cx_2).$$

The Lyapunov function is positive-definite since the only term that causes it to differ from a function for a linear differential equation is non-negative. It is interesting to consider the case where the linear system has negative damping. For simplicity, let $K=\omega=c=1$, and let $D=-\frac{1}{2}$. Then

$$V = \frac{3}{2}{x_1}^2 - x_1x_2 + (x_1 + x_2)\operatorname{sgn}(x_1 + x_2) + {x_2}^2,$$

and

$$\dot{V} = {x_1}^2 + {x_2}^2 - 1.$$

The derivative is negative inside of a unit circle which is shown by the dashed curve in Fig. 8. The curve $V=7/4$

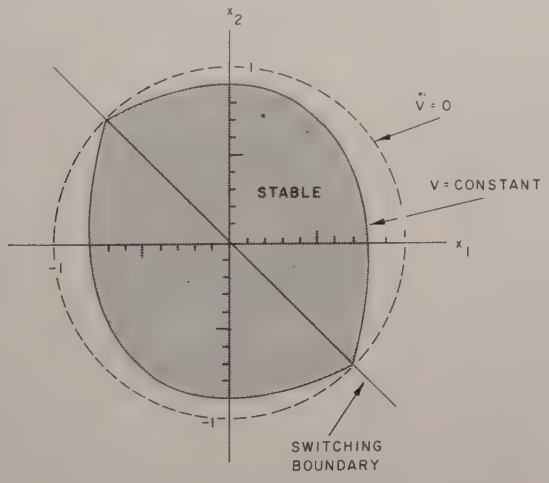


Fig. 8—Stable region for negative damping.

touches this circle at the switching boundaries and, according to the modified theorem given in Section III, encloses a stable region. It may be concluded that this is a good approximation since the curves, $V=$a constant and $\dot{V}=0$, fall so close together.

A General Nonlinearity

The third-order equation

$$\dddot{y} + a_1\ddot{y} + a_2\dot{y} + f(y)y = 0$$

is asymptotically B_∞ stable if the roots of its characteristic equation, calculated as if $f(y)$ were constant, have negative-real parts. To prove this, a matrix is derived from (10) by letting $\mathbf{C}(x)$ be the following matrix:

$$\begin{pmatrix} 2a_2(yf' + f) & 0 & 2(yf' + f) \\ 0 & 0 & 0 \\ 2(yf' + f) & 0 & 2a_1 \end{pmatrix}.$$

$\boldsymbol{B}(\mathbf{x})$ has the form (11), and the resulting matrix is

$$\boldsymbol{A}(\mathbf{x}) = \begin{pmatrix} {a_2}^2 + a_1(yf' + f) & a_1a_2 & a_2 \\ a_1a_2 & {a_1}^2 + a_2 & a_1 \\ a_2 & a_1 & 2 \end{pmatrix}.$$

Integrating this matrix gives the three quantities:

$$\nabla V = \begin{pmatrix} {a_2}^2y + a_1f(y)y + a_1a_2\dot{y} + a_2\ddot{y} \\ a_1a_2y + ({a_1}^2 + a_2)\dot{y} + a_1\ddot{y} \\ a_2y + a_1\dot{y} + 2\ddot{y} \end{pmatrix},$$

$$V = \frac{{a_2}^2}{2}y^2 + a_1\int_0^y f(y)y\,dy + a_1a_2y\dot{y} + ({a_1}^2 + a_2)\frac{\dot{y}^2}{2} + a_2y\ddot{y} + a_1\dot{y}\ddot{y} + \ddot{y}^2,$$

and

$$V = [a_2f(y)y^2 + 2f(y)y\ddot{y} + a_1\ddot{y}^2].$$

The derivative is nonpositive under the conditions $a_1>0$, $a_2>0$, $f(y)>0$, and $a_1a_2-f(y)>0$. The Lyapunov function is positive definite if the first three are fulfilled. These conditions are just the Routh-Hurwitz inequalities for the characteristic equation defined above and are necessary and sufficient for the characteristic roots to have negative-real parts, which proves the statement. It cannot be concluded that these are necessary requirements for stability since if they are not satisfied the derivative becomes indefinite but not positive semi-definite.

Conclusions

A criterion for stability and a method for applying it to many of the nonlinear problems associated with control systems have been presented. The approach has been to admit that a certain amount of experience and ingenuity are required if other than very specialized situations are to be handled, but a plan of attack is developed which quickly leads to the correct answers to some questions. The procedure is designed to permit a number of modifications for treating cases that are not amenable to the categorical application of the process. The examples were selected to illustrate various points concerned with application and modification of the method rather than for being representative of practical problems.

Because the criterion gives only sufficient conditions, some judgment is required in evaluating the results. No fixed rules for doing this can be set down but certain

guides are possible. If the same requirements for which the derivative is nonpositive are also those which cause the Lyapunov function to be positive definite or monotonically increasing, the result is usually very good. In other words, both the Lyapunov function and its derivative should satisfy the conditions imposed upon them with the least possible margin. If one is maximized at the expense of the other, an inferior result is obtained.

Modifications of the functions derived by direct application of the method are made with the objective of approaching this condition. It is often evident that the addition or deletion of a particular term will produce this effect, in which case the variation should be made. When both requirements are fulfilled by a wide margin, there are many adequate Lyapunov functions, and there is no need to seek an optimum.

This discussion has been concerned entirely with the generation of Lyapunov functions for stability analysis. As with other methods, stability analysis by the use of Lyapunov functions is closely related to other facets of control system design. Hahn [4] and Kalman and Bertram [5] describe some uses of the Lyapunov method for other performance considerations. The functions derived by the method presented here are suitable for use in these applications as well.

Acknowledgment

The author wishes to thank Dr. I. Flügge-Lotz of Stanford University, Stanford Calif., under whose guidance the dissertation upon which this paper is based was prepared.

Bibliography

[1] D. R. Ingwerson, "A Modified Lyapunov Method for Nonlinear Stability Problems," Ph.D. dissertation, Stanford University, Stanford, Calif.; November, 1960.
[2] D. R. Ingwerson, "Principal Definitions of Stability," AIEE Workshop on Lyapunov's Second Method, September, 1960. Published by the University of Michigan Industry Program of the College of Engineering, Ann Arbor, Mich.; September, 1960.
[3] A. M. Lyapunov, "Problème général de la stabilité du mouvement," Princeton University Press, Princeton, N. J.; 1949. Reprinted in French from original translation.
[4] W. Hahn, "Theorie und Anwendung der Direkten Methode von Ljapunov," Springer Verlag, Berlin, Germany; 1959.
[5] R. E. Kalman and J. E. Bertram, "Control system analysis and design via the 'Second Method' of Lyapunov," *ASME J. of Basic Engrg.*, Paper No. 59—NAC-2; November, 1959.
[6] W. J. Cunningham, "An Introduction to Lyapunov's Second Method," AIEE Workshop on Lyapunov's Second Method, September, 1960. Published by the University of Michigan Industry Program of the College of Engineering, Ann Arbor, Mich.; September, 1960.
[7] E. V. Barbasin and N. N. Krasovskii, "On stability of motion in the large," *Dok. Akad. Nauk.*, vol. 86, pp. 454–456; 1952.
[8] N. N. Krasovskii, "On stability with large initial disturbances," *Prikladnaya Matematika i Mekhanika*, vol. 21, pp. 309–319; 1957.
[9] H. Antosiewicz, "A Survey of Lyapunov's Second Method," *Ann. Mathematics Studies*, No. 41, Princeton University Press, Princeton, N. J., pp. 147–166; 1959.
[10] N. N. Krasovskii, "On application of the second method of A. M. Liapunov to equations with time delays," *Prikladnaya Matematika i Mekhanika*, vol. 20, pp. 315–327; 1956. (In Russian.)
[11] W. Hahn, "Eine Bemerkung zur zweiten Methode von Ljapunov," *Math. Nachr.*, vol. 14, pp. 349–354; 1956.
[12] N. N. Krasovskii, "On the stability in the large of a system of nonlinear differential equations," *Prikladnaya Matematika i Mekhanika*, vol. 18, pp. 735–737; 1954. (In Russian.)
[13] E. V. Barbasin, "On the stability of solutions of a third order nonlinear differential equation," *Prikladnaya Mathematika i Mekhanika*, vol. 16, p. 629; 1952. (In Russian.)
[14] M. A. Aizerman, "On a problem concerning the in-the-large stability of dynamic systems," *Uspekhi Matematika Nauk*, vol. 4, pp. 187–188; 1949. (In Russian.)

A Mean-Weighted Square-Error Criterion for Optimum Filtering of Nonstationary Random Processes*

G. J. MURPHY†, MEMBER, IRE, AND K. SAHARA†

Summary—A procedure for use in the design of a physically realizable time-invariant linear system for optimum filtering of a nonstationary random process in the presence of nonstationary random noise is presented in this paper. First, a new criterion for system performance is defined. On the basis of this criterion, an integral equation for the optimum physically realizable weighting function is derived, and it is shown that in some cases an exact solution to this equation can be obtained through the use of double Fourier transforms. Then the use of a technique to obtain an approximation to the solution to the integral equation is discussed.

This theoretical background is followed by an illustrative example in which the method is used to design the optimum physically realizable linear time-invariant filter for a Brownian-motion signal contaminated by Markovian noise. It is shown here that if the designer is constrained by the requirement that the system be a digital filter with finite memory, then an exact solution can be found.

Application of the method in cases where the random processes are stationary is discussed next, and the suggested approach is illustrated in an example.

Introduction

THE EARLY work of Wiener[1] on optimum synthesis of a time-invariant linear physically realizable system has been extended[2–8] in several respects during the past two decades. Another extension is suggested in this paper.

In some of the work[5] noted above, the criterion for optimum performance is changed by changing the weighting of the error as a function of the error magnitude. In other work,[7,8] the criterion is modified by changing the weighting of the error as a function of the time at which the error occurs. It is proposed in this paper that a still different criterion be employed—a criterion in which the importance of the error is made a function of a subset of the sample space from which the error is taken. In the simplest application of this kind of error weighting, the criterion is of the form

$$C = \sum_j m_j \sum_k P_{jk} e_{jk}^2(t), \tag{1}$$

if there is a finite or a countably infinite number of subsets and sample points in each subset, and of the form

$$C = \int_{-\infty}^{\infty} m(x) \int_{-\infty}^{\infty} p(x, y) e^2(x, y; t) dy dx, \tag{2}$$

if there is an uncountably infinite number of subsets and of sample points in each subset. In (1), P_{jk} is the probability associated with the kth sample point in the jth subset, and m_j is a subset weighting parameter. In (2), $p(x, y)$ is a probability density function and $m(x)$ is a subset weighting function. Clearly, if $m_j \equiv 1$ in (1) and $m(x) \equiv 1$ in (2), then this criterion reduces to the familiar mean-square error criterion. In general, however, it is a weighted mean-square error criterion.

As an example of a possible application of this criterion, consider the problem of designing the suspension system of an automobile for the "best ride." One might obtain extensive data on road conditions and proceed to design the system by use of the conventional mean-square error criterion. It seems reasonable, however, to subdivide road conditions into the categories of city streets, country roads, and expressways (*i.e.*, freeways, tollways, turnpikes). It is to be expected that the statistical characteristics of these three categories differ from each other and from those of the complete ensemble, as well. It is now possible to introduce weighting parameters m_1, m_2, and m_3 that indicate the relative importance of a "smooth ride" in the city, on the country road, and on an expressway, respectively, as well as the probabilities P_1, P_2, and P_3 that the automobile will be on these respective kinds of roads. Then the criterion of performance can be taken to be

$$C = m_1 P_1 \underset{\sim}{e_1^2}(t) + m_2 P_2 \underset{\sim}{e_2^2}(t) + m_3 P_3 \underset{\sim}{e_3^2}(t), \tag{3}$$

where $\underset{\sim}{e_j^2}(t)$ denotes the ensemble average of $e^2(t)$ over the jth subset. Inasmuch as it would be impractical to provide the automobile with three separate suspension systems that could be switched into or out of use depending on the type of road being traversed, a single

* Received by the PGAC, December 20, 1960.

† Elec. Engrg. Dept., Northwestern University, Evanston, Ill.

[1] N. Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series," John Wiley and Sons, Inc., New York, N. Y.; 1949.

[2] R. C. Booton, Jr., "An optimization theory for time-varying linear systems with non-stationary statistical inputs," PROC. IRE, vol. 40, pp. 977–981; August, 1952.

[3] L. A. Zadeh and J. R. Ragazzini, "An extension of Wiener's theory of prediction," *J. Appl. Phys.*, vol. 21, pp. 645–655; July, 1950.

[4] M. G. Spooner and V. C. Rideout, "Correlation studies of linear and nonlinear systems," *Proc. NEC*, Chicago, Ill., October 1–3, 1956, vol. 12, pp. 321–335; 1956.

[5] W. C. Schultz and V. C. Rideout, "A general criterion for servo performance," *Proc. NEC*, Chicago, Ill., October 1–7, 1957, vol. 13, pp. 549–560; 1957.

[6] M. V. Mathews and C. W. Steeg, "Final value controller synthesis," IRE TRANS. ON AUTOMATIC CONTROL, vol. 2, pp. 6–16; February, 1957.

[7] J. Zaborszky and J. W. Diesel, "Probabilistic error as a measure of control system performance," *Trans. AIEE* (Application and Industry), pp. 163–168; July, 1959.

[8] G. J. Murphy and N. T. Bold, "Optimization based on a square-error criterion with an arbitrary weighting function," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 24–30; January, 1960.

suspension system that is optimized with respect to the criterion in (3) appears to be the best system that can be constructed.

Similarly, it would seem reasonable to design a fire-control system on the basis of a weighted mean-square error criterion, with a different weighting factor for bombers than for fighters. The weighting factors could be chosen on the basis of the destructive potentials and the relative target size, for example, of these two classes of aircraft.

Furthermore, particularly in those instances in which one or more of the random processes is nonstationary, it is sometimes desirable to weight the error as a function of time in addition to weighting it according to the subset from which the signal is drawn. (One may wish, for example, to accentuate the importance of error at a set of sampling instants.) This can be accomplished by the use of an integral of weighted square error or by the use of a time-mean weighted square error. Thus, the criteria presented in (1) and (2) can be generalized to

$$C = \sum_j m_j \sum_k P_{jk} \int_{-\infty}^{\infty} w_{jk}(t) e_{jk}^2(t)\,dt, \tag{1a}$$

or

$$C = \sum_j m_j \sum_k P_{jk} \lim_{T_o \to \infty} \frac{1}{2T_o} \int_{-T_o}^{T_o} w_{jk}(t) e_{jk}^2(t)\,dt, \tag{1b}$$

and

$$C = \int_{-\infty}^{\infty} m(x) \int_{-\infty}^{\infty} p(x, y) \int_{-\infty}^{\infty} w(x, y; t) e^2(x, y; t)\,dt\,dy\,dx, \tag{2a}$$

or

$$C = \int_{-\infty}^{\infty} m(x) \int_{-\infty}^{\infty} p(x, y) \lim_{T_o \to \infty} \frac{1}{2T_o} \cdot \int_{-T_o}^{T_o} w(x, y; t) e^2(x, y; t)\,dt\,dy\,dx. \tag{2b}$$

It should be noted that if $w_{jk}(t) = w(x, y; t) = u_0(t)$, where $u_0(t)$ is a unit impulse function at $t = 0$, then (1a) and (2a) reduce to (1) and (2), respectively. Similarly, albeit somewhat artificially, (1b) and (2b) can be reduced to (1) and (2), respectively, by letting $w_{jk}(t) = w(x, y; t) = 2T_o u_0(t)$. For sampled-data applications, the weighting function $w_{jk}(t)$ can be taken to be $w_{jk} i(t)$, where $i(t)$ is the unit impulse train, to obtain the sum of the weighted squares of the error samples; and for $w_{jk} \equiv 1$, this reduces to the commonly used sum of the squares of the error samples.

Arbitrary magnitude weighting of the error can be introduced, in addition to the time weighting and set weighting indicated in (1a), (1b), (2a), and (2b), by replacing $e^2(t)$ with $F[e(t)]$. Thus, the general form of the criterion becomes

$$C \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \lim_{T_o \to \infty} \int_{-T_o}^{T_o} w(x, y; T_o; t) \cdot F[e(x, y; t)]\,dt\,dy\,dx \tag{4}[9]$$

$$= \left\langle \left\langle \lim_{T_o \to \infty} \int_{-T_o}^{T_o} w(x, y; T_o; t) F[e(x, y; t)]\,dt \right\rangle_y \right\rangle_x \tag{5}$$

where $\langle\ \rangle_x$ and $\langle\ \rangle_y$ denote averages taken over x and y, respectively.

This criterion is hereafter referred to as the "MASTWE" (Mean-Amplitude-Set-Time Weighted Error) criterion.

Optimum Synthesis

For the present,[10] let

$$F[e(x, y; t)] = e^2(x, y; t). \tag{6}$$

Then, (5) becomes

$$C = \left\langle \left\langle \lim_{T_o \to \infty} \int_{-T_o}^{T_o} w(x, y; T_o; t) e^2(x, y; t)\,dt \right\rangle_y \right\rangle_x \tag{7}$$

$$= \lim_{T_o \to \infty} \int_{-T_o}^{T_o} \langle \langle w(x, y; T_o; t) e^2(x, y; t) \rangle_y \rangle_x dt \tag{8}$$

if the order of the various operations may be changed as indicated.[11]

Following the customary procedure, it is convenient now to define the system error $e(t)$ as the difference between an ideal response $i(t)$ and the actual response $c(t)$. That is,

$$e(t) \triangleq i(t) - c(t). \tag{9}$$

The relation between the actual response and the input to the system is determined by the weighting function $g(\tau)$, if the system is a linear constant-parameter system. For such a system,

$$c(t) = \int_{-\infty}^{\infty} g(\tau) r(t - \tau)\,d\tau. \tag{10}$$

Eqs. (9) and (10) are presented graphically in Fig. 1.

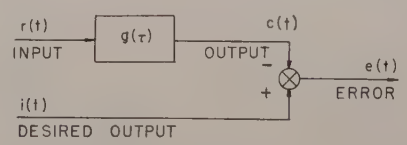


Fig. 1—A diagrammatic definition of system error.

[9] It should be noted that the impulse functions of x, y, and t may be contained in $p(x, y)$ and $w(x, y; T_o; t)$.

[10] Other forms of $F[e(x, y; t)]$ are to be investigated and reported in future papers.

[11] For a discussion of conditions under which this change in order of operations is justifiable, see, for example, E. W. Hobson, "The Theory of Functions of a Real Variable," Harren Press, Washington, D. C., 2 vols.; 1950.

To simplify the notation, the following definitions are now introduced:

$$\psi_{wii}(\tau_1, \tau_2) \triangleq \lim_{T_o \to \infty} \int_{-T_o}^{T_o} \langle\langle w(x, y; T_o; t) i(x, y; t + \tau_1) \cdot i(x, y; t + \tau_2)\rangle_y\rangle_x dt, \quad (11)$$

$$\psi_{wri}(\tau_1, \tau_2) \triangleq \lim_{T_o \to \infty} \int_{-T_o}^{T_o} \langle\langle w(x, y; T_o; t) r(x, y; t + \tau_1) \cdot i(x, y; t + \tau_2)\rangle_y\rangle_x dt, \quad (12)$$

$$\psi_{wrr}(\tau_1, \tau_2) \triangleq \lim_{T_o \to \infty} \int_{-T_o}^{T_o} \langle\langle w(x, y; T_o; t) r(x, y; t + \tau_1) \cdot r(x, y; t + \tau_2)\rangle_y\rangle_x dt. \quad (13)$$

Each of these three functions is a second-order correlation function.

By substitution of (9) and (10) into (8) and introduction of (11), (12), and (13), it can be shown that

$$C = \psi_{wii}(0, 0) - 2\int_{-\infty}^{\infty} g(\tau)\psi_{wri}(-\tau, 0) d\tau + \int_{-\infty}^{\infty} g(\tau_1) \int_{-\infty}^{\infty} g(\tau_2)\psi_{wrr}(-\tau_1, -\tau_2) d\tau_2 d\tau_1. \quad (14)$$

The optimum weighting function $g_0(\tau)$, for which C assumes its minimum value, is now to be determined.

Following the customary procedure based on the calculus of variations, we now let

$$g(\tau) = g_o(\tau) + \epsilon h(\tau), \quad (15)$$

where $h(\tau)$ is an arbitrary physically realizable weighting function satisfying the relation

$$h(\tau) \equiv 0, \qquad \tau \leq 0, \qquad \tau > T_M, \quad (16)$$

where T_M is the length of memory of the system, with no discontinuity at $\tau = 0$ or at $\tau = T_M$, and where ϵ is an arbitrary real parameter. A necessary condition for an optimum solution is

$$\frac{\partial C}{\partial \epsilon} = 0, \qquad \epsilon = 0. \quad (17)$$

Thus, it is found that the optimum weighting function must satisfy the relation

$$\int_{-\infty}^{\infty} h(\tau_1) \left[\psi_{wri}(-\tau_1, 0) - \int_{-\infty}^{\infty} g_o(\tau_2)\psi_{wrr}(-\tau_1 - \tau_2) d\tau_2 \right] d\tau_1 = 0, \quad (18)$$

if the usual change in the order of carrying out the various operations can be justified.

Since $h(\tau)$ is arbitrary and satisfies (16), (18) is equivalent to the requirement that

$$\int_{-\infty}^{T_M} g_o(\tau_2)\psi_{wrr}(-\tau_1, -\tau_2) d\tau_2 = \psi_{wri}(-\tau_1, 0), \qquad 0 \leq \tau_1 < T_M. \quad (19)$$

Whether satisfaction of (19) is also a sufficient condition to insure that the $g(\tau)$ in (15) yields a minimum of $C(\epsilon)$ for $\epsilon = 0$ can be ascertained by examining

$$\frac{\partial^2 C}{\partial \epsilon^2} = 2\int_{-\infty}^{\infty} h(\tau_1) \int_{-\infty}^{\infty} h(\tau_2)\psi_{wrr}(-\tau_1, -\tau_2) d\tau_2 d\tau_1. \quad (20)$$

If the change in order of the operations can be justified, (20) may be rewritten as

$$\frac{\partial^2 C}{\partial \epsilon^2} = 2 \lim_{T_o \to \infty} \int_{-T_o}^{T_o} \langle\langle w(x, y; T_o; t) c_h^2(x, y; t)\rangle_y\rangle_x dt, \quad (21)$$

where $c_h(x, y; t)$ is the response of a system with weighting function $h(\tau)$ at time t to the random input $r(t)$. Clearly, $(\partial^2 C/\partial \epsilon^2) > 0$ if $w(x, y; T_o; t) > 0$ for all x, y, T_o, and t.

A sufficient condition that $g_0(\tau)$ satisfy (19) is that

$$G_o(j\omega) = \frac{\Phi_{wir}(j\omega, j\omega')}{\Phi_{wrr}(j\omega, j\omega')}, \quad (19a)$$

where $G_0(j\omega)$ is a Fourier transform of $g_0(\tau)$ and $\Phi_{wir}(j\omega, j\omega')$ and $\Phi_{wrr}(j\omega, j\omega')$ are double Fourier transforms of $\psi_{wir}(\tau, \gamma)$ and $\psi_{wrr}(\tau, \gamma)$, respectively. However, in some cases these double Fourier transforms do not exist, and in other cases the right-hand side of (19a) is not independent of ω'. It is therefore not always possible to obtain an exact solution that satisfies the sufficient condition (19a).

An approximation to the solution to (19) can be obtained by regarding $g_0(\tau)$ as a staircase function of τ (instead of a smooth function) and then requiring that (19) be satisfied only for a finite set of discrete values of $\tau_1 > 0$. That is, (19) is replaced with

$$T \sum_{j=-L}^{N} g_o(jT)\psi_{wrr}(-kT, -jT) = \psi_{wri}(-kT, 0), \qquad k = 0, 1, 2, \cdots, N, \quad (22)$$

where T is an arbitrarily small positive real number, and L and N are determined by the length of memory anticipated in the optimum system. Furthermore, the optimum physically realizable weighting function $g_{or}(t)$ is approximated by the solution to

$$T \sum_{j=0}^{N} g_{or}(jT)\psi_{wrr}(-kT, -jT) = \psi_{wri}(-kT, 0), \qquad k = 0, 1, 2, \cdots, N. \quad (23)$$

Finally, (23) may be written as

$$\mathbf{\Psi}_{wrr}\, \mathbf{G}_{or} = \frac{1}{T}\, \mathbf{\Psi}_{wri}, \quad (24)$$

where

$$G_{or} \triangleq \begin{bmatrix} g_{or}(0) \\ g_{or}(T) \\ \vdots \\ g_{or}(NT) \end{bmatrix}, \tag{25}$$

$$\Psi_{wrr} \triangleq \begin{bmatrix} \psi_{wrr}(0, 0) & \psi_{wrr}(0, -T) & \cdots & \psi_{wrr}(0, -NT) \\ \psi_{wrr}(-T, 0) & \psi_{wrr}(-T, -T) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \psi_{wrr}(-NT, 0) & \psi_{wrr}(-NT, -T) & \cdots & \psi_{wrr}(-NT, -NT) \end{bmatrix}, \tag{26}$$

and

$$\Psi_{wri} \triangleq \begin{bmatrix} \psi_{wri}(0, 0) \\ \psi_{wri}(-T, 0) \\ \vdots \\ \psi_{wri}(-NT, 0) \end{bmatrix}. \tag{27}$$

If a solution to (24) exists, it can be found by the usual method. The accuracy of the approximation is, of course, a function of N and T. The designer will naturally choose T as small as possible and N as large as possible, subject to the limitations of the available data and the computer to be used in solving (24) for G_{or}.

When a solution to (24) has been obtained, the design can be completed by fitting the curve of $g_{or}(kT)$ vs kT with exponentials and exponentially decaying sinusoids and then designing a realizable device with unit impulse response equal to the sum of these components. It should be noted that it is no more difficult to design a digital system (with sampling period T) than to design a continuous system by this method. Furthermore, if the system is to be a digital system with finite memory, an exact solution can be obtained, as is illustrated in the following example.

Example 1

To illustrate the use of the MASTWE criterion, the following example is presented:

Problem: The input $r(t)$ to the system illustrated in Fig. 1 is

$$r(t) = s(t) + n(t), \tag{28}$$

where $s(t)$ is a Brownian motion with autocorrelation function

$$\phi_{ss}(t - x, t - y) = \begin{cases} 25 \min (t - x, t - y), \min (t - x, t - y) \geq 0 \\ 0, \min (t - x, t - y) < 0 \end{cases}, \tag{29}$$

and $n(t)$ is a stationary random noise with autocorrelation function

$$\phi_{nn}(t - x, t - y) = 16e^{-2|x-y|}. \tag{30}$$

The system is to be a digital (or sampled-data) system operating on samples taken at $t = 0, 1, 2, \cdots$, with finite memory of length M, where M is a positive integer.

The two functions $s(t)$ and $n(t)$ are independent, and it is desired that the output equal $s(t)$ at the sampling instants. The weighting function to be used, which is

$$w(x, y; T_o; t) = te^{-t} \sum_{n=0}^{\infty} u_o(t - n), \tag{31}$$

is chosen to accentuate the importance of minimizing the error at the sampling instants in general, and at $t = 1$, in particular.

Determine the optimum physically realizable weighting function.

Solution: First, the second-order correlation functions are determined. These are

$$\psi_{wrr}(-k, -j) = \lim_{T \to \infty} \int_{-T_o}^{T_o} te^{-t} \sum_{n=0}^{\infty} u_o(t - n)\phi_{rr}(t - k, t - j)dt \tag{32}$$

$$= \sum_{n=0}^{\infty} ne^{-n}\phi_{rr}(n - k, n - j) \tag{33}$$

$$= \begin{cases} 25 \sum_{n=k}^{\infty} ne^{-n}(n - k) + 16 \sum_{n=0}^{\infty} ne^{-n}e^{-2|k-j|}, k \geq j \geq 0 \\ 25 \sum_{n=j}^{\infty} ne^{-n}(n - j) + 16 \sum_{n=0}^{\infty} ne^{-n}e^{-2|k-j|}, 0 \leq k \leq j \end{cases} \tag{34}$$

and

$$\psi_{wri}(-k, 0) = 25 \sum_{n=k}^{\infty} ne^{-n}(n - k), \quad k \geq 0. \tag{35}$$

Numerical values of these functions, obtained by addition of up to the first fifteen terms in the series, are presented in Tables I and II.

TABLE I
SECOND-ORDER CROSS-CORRELATION FUNCTIONS

k	$\psi_{wri}(-k, 0)$
0	49.8
1	26.7
2	12.8
3	5.9

TABLE II
SECOND-ORDER AUTOCORRELATION FUNCTIONS
$\psi_{wrr}(-k, -j)$

k \ j	0	1	2	3
0	64.5	28.7	13.1	5.9
1	28.7	41.4	14.7	6.1
2	13.1	14.7	27.5	7.9
3	5.9	6.1	7.9	20.6

By choosing a length of memory (M) from 0 through 3 and then substituting into (26) and (27) the appropriate values from Tables I and II and solving[12] (24), one can now determine the optimum physically realizable weighting sequence $g(n)$, $n=0, 1, \cdots, M$. The results are presented in Table III.

TABLE III
OPTIMUM REALIZABLE WEIGHTING SEQUENCE

M	$g(0)$	$g(1)$	$g(2)$	$g(3)$
0	0.77	0	0	0
1	0.70	0.16	0	0
2	0.70	0.14	0.08	0
3	0.70	0.14	0.06	0.02

APPLICATIONS TO STATIONARY RANDOM PROCESSES

If the random processes under consideration are stationary and

$$w(x, y; T_o; t) = w_1(x, y)w_2(T_o, t), \tag{36}$$

then (8) may be rewritten as

$$C = W\langle\langle w_1(x, y)e^2(x, y; t)\rangle_y\rangle_x, \tag{37}$$

where

$$W \triangleq \lim_{T_o\to\infty} \int_{-T_o}^{T_o} w_2(T_o, t)dt. \tag{38}$$

Thus, it is seen that in such a case the optimum weighting function is independent of the choice of $w_2(T_o, t)$.

Now let

$$\delta_{ri}(\tau_1 - \tau_2) \triangleq \langle\langle w_1(x, y)r(t - \tau_1)i(t - \tau_2)\rangle_y\rangle_x \tag{39}$$

and

$$\delta_{rr}(\tau_1 - \tau_2) \triangleq \langle\langle w_1(x, y)r(t - \tau_1)r(t - \tau_2)\rangle_y\rangle_x. \tag{40}$$

Then it can be shown that for the case under consideration (19) becomes

$$\int_{-\infty}^{T_M} g_o(\tau_2)\delta_{rr}(\tau_1 - \tau_2)d\tau_2 = \delta_{ri}(\tau_1), \quad 0 \le \tau_1 < T_M. \tag{41}$$

If the memory is of infinite length (*i.e.*, $T_M = \infty$), then (41) is the Wiener-Hopf integral equation, and the exact solution for the optimum physically realizable transfer function is known to be

$$G_{or}(s) = \frac{\left[\dfrac{\Delta_{ri}(s)}{\Delta_{rr}^{-}(s)}\right]_+}{\Delta_{rr}^{+}(s)}, \tag{42}$$

where $\Delta_{ri}(s)$ and $\Delta_{rr}(s)$ are bilateral Laplace transforms of $\delta_{ri}(\tau)$ and $\delta_{rr}(\tau)$, respectively, and $\Delta_{rr}^{-}(s)$ and $\Delta_{rr}^{+}(s)$ are obtained in the customary manner by spectrum factorization of $\Delta_{rr}(s)$, and

$$\left[\frac{\Delta_{ri}(s)}{\Delta_{rr}^{-}(s)}\right]_+ \triangleq \int_0^\infty e^{-s\tau}\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} e^{s\tau}\frac{\Delta_{ri}(s)}{\Delta_{rr}^{-}(s)}\,ds d\tau. \tag{43}$$

Example 2

Problem: The input $r(t)$ to the system illustrated in Fig. 1 is equally likely to be

$$u(t) = s(t) + n(t) \tag{44}$$

or

$$v(t) = d(t) + n(t), \tag{45}$$

where $s(t)$ and $d(t)$ are signals and $n(t)$ is noise, and all three are generated by stationary random processes. The desired output $i(t)$ is the signal component of the input. The power spectral densities of $s(t)$, $d(t)$, and $n(t)$ are

$$\Phi_{ss}(s) = \frac{9}{9 - s^2}, \tag{46}$$

$$\Phi_{dd}(s) = \frac{4}{4 - s^2}, \tag{47}$$

and

$$\Phi_{nn}(s) = 2, \tag{48}$$

respectively, and $n(t)$ is uncorrelated with both $s(t)$ and $d(t)$.

Determine the optimum physically realizable transfer function for the system, given that it is twice as important to process $d(t)$ accurately as it is to process $s(t)$ accurately.

[12] That the solution thus obtained does indeed yield a minimum of C for $g(\tau)$ described by (15) can easily be ascertained by noting that since $w(x, y; T_o; t)$ is a non-negative function, $(\partial^2 C/\partial e^2) > 0$.

Solution: From the given information it is found that

$$\Delta_{rr}(s) = \frac{1}{2}\left(\frac{9}{9-s^2}+2\right)+\frac{1}{2}\times 2\times\left(\frac{4}{4-s^2}+2\right)$$

$$= \frac{3(s^2-10.87)(s^2-4.97)}{(s^2-9)(s^2-4)} \tag{49}$$

and

$$\Delta_{ri}(s) = \frac{1}{2}\left(\frac{9}{9-s^2}\right)+\frac{1}{2}\times 2\times\left(\frac{4}{4-s^2}\right)$$

$$= -\frac{17}{2}\,\frac{\left(s^2-\frac{108}{17}\right)}{(s^2-9)(s^2-4)}. \tag{50}$$

It follows that

$$\Delta_{rr}^{+}(s) = \sqrt{3}\,\frac{(s+3.3)(s+2.2)}{(s+3)(s+2)} \tag{51}$$

and

$$\Delta_{rr}^{-}(s) = \sqrt{3}\,\frac{(s-3.3)(s-2.2)}{(s-3)(s-2)}, \tag{52}$$

that

$$\frac{\Delta_{ri}(s)}{\Delta_{rr}^{-}(s)}$$

$$= -\frac{17}{2\sqrt{3}}\,\frac{s^2-\frac{108}{17}}{(s+3)(s+2)(s-3.3)(s-2.2)} \tag{53}$$

$$= \frac{1}{2\sqrt{3}}\left(\frac{1.37}{s+3}+\frac{1.79}{s+2}-\frac{2.13}{s-3.3}-\frac{0.99}{s-2.2}\right), \tag{54}$$

and hence that

$$\left(\frac{\Delta_{ri}(s)}{\Delta_{rr}^{-}(s)}\right)_{+} = \frac{3.16s+8.11}{2\sqrt{3}\,(s+3)(s+2)}. \tag{55}$$

Finally, by substituting (55) and (51) into (42), it is found that

$$G_{or}(s) = 0.53\,\frac{s+2.55}{(s+2.2)(s+3.3)}. \tag{56}$$

It is interesting to note that if the probability of $u(t)$ and $v(t)$ were 1 and 0, respectively, the optimum realizable transfer function would be

$$G_{or}(s) = \frac{0.67}{s+3.67}, \tag{57}$$

that if the probabilities of $u(t)$ and $v(t)$ were 0 and 1, respectively, the optimum realizable transfer function would be

$$G_{or}(s) = \frac{0.45}{s+2.45}, \tag{58}$$

and that if $u(t)$ and $v(t)$ were equally likely and equally important, the optimum realizable transfer function would be

$$G_{or}(s) = 0.563\,\frac{s+2.39}{(s+2.16)(s+3.4)}. \tag{59}$$

Conclusions

The mean-square error criterion has been generalized to obtain the mean amplitude-set-time-weighted error criterion. By suitable choice of the weighting functions in this criterion, one can accentuate the importance of error of specified amplitudes, of error occurring at specified times, and of error in processing specified inputs. The use of this criterion leads to the synthesis of the optimum time-invariant physically realizable system for performing the desired task.

Although the inputs to the system may be either stationary or nonstationary random processes, this criterion is particularly useful for treatment of nonstationary random processes.

An exact solution for the optimum realizable transfer function is not obtainable in general. However, in certain special cases the exact solution can readily be found. Notable among these special cases are those in which 1) square of error is used for amplitude weighting and 2) the system is a digital system with finite memory, or the inputs are stationary and the set and time weighting functions are separable. For many cases in which an exact solution cannot be found, an approximation can be obtained by the use of familiar numerical methods.

The success of synthesis on the basis of the MASTWE criterion in the special cases mentioned above is encouraging. Possibilities for future work on this subject include consideration of other amplitude-weighting functions, of rules for choosing the form of the set-time weighting function for specific applications, and of the properties of the second-order correlation functions and their transforms.

A General Performance Index for Analytical Design of Control Systems*

Z. V. REKASIUS†

Summary—A generally applicable performance index for linear feedback systems is proposed. This performance index enables one to specify a desired (ideal) response towards which the system is optimized. The response of any unity-numerator system can be chosen as the ideal response. The coefficients of the performance index are determined from the specified ideal response, and the procedure of optimization is equivalent to minimization of the numerical value of this performance index. The procedure for constructing Liapunov functions for linear systems is used to minimize the performance index. System optimization is illustrated by means of examples. The results indicate that the smaller the numerical value of the performance index, the closer the actual system response approaches the specified ideal response.

Introduction

SEVERAL performance indexes have been proposed in the literature for use in connection with the design and optimization of linear automatic control systems.[1,2]

The ISE performance index, defined as

$$I = \int_0^\infty e^2(t)dt \qquad (1)$$

has been used, for example, by Newton, Gould and Kaizer[3] in the analytical design of optimum systems. The ISE performance index yields systems that contain long sustained oscillations. For example, in a second-order system the optimum ISE system was found to require infinite loop gain and resulted in decaying oscillations of infinite frequency—certainly a very undesirable situation. The basic weakness of the ISE and other similar error-integral type performance indexes is that the response of the optimum systems does not follow any predictable form; *i.e.*, the designer does not have any *a priori* knowledge about the desirable (or undesirable) performance characteristics of the optimum system.

In order that a performance index be generally applicable, it not only must reveal the performance characteristics of its optimum systems, but also must enable the designer to choose what the desired characteristics of the optimum system should be. In some applications, such as line voltage regulation, for example, an extremely fast response to a step disturbance may be desirable. In other applications, such as control of an airplane in flight, too fast a response to a step command may be unacceptable.

The automatic control system frequently represents only a part (*i.e.*, a subsystem) of an over-all physical system (missile, airplane, chemical process, etc.). The performance requirements for the control system are thus determined by the mission of the over-all system, and are not the choice of the designer of the control subsystem. Consequently, the desired performance specifications of control systems are expressed either graphically, as desired response curves, or in the form of differential equations and not in the terms of control-system design criteria (such as damping, phase margin, M-peak, bandwidth, etc.).

The performance index proposed in this paper enables the designer to select the ideal transient response which could be achieved if a sufficient number of system parameters were under the designer's control (*i.e.*, a free-system configuration). This ideal response, expressed by a linear homogeneous differential equation, corresponds to the absolute minimum value $I_{\min\ \min}$ of the performance index I. In the practical case of semi-free-system configuration, the variable parameters are adjusted to yield the minimum value $I_{\min}$ of this performance index. The smaller the difference between the values of the performance index of the ideal model $I_{\min\ \min}$ and the optimum system $I_{\min}$, the closer the transient response of the optimum system will approach the specified ideal response. Hence, the particular performance index represents a different philosophy of the analytical design. The analytical design procedure based upon the proposed performance index does not attempt to yield a "good" response, but rather it optimizes towards a specific transient response selected to suit the needs of a particular application of the system. The ability to incorporate the desired transient response into this performance index makes it a general performance index for the analytical design of linear control systems.

A simple straightforward procedure of calculating this performance index is outlined in the paper. This procedure consists of a solution for a set of linear algebraic equations. Simultaneous solution of these algebraic equations yields the value of performance index in terms

* Received by the PGAC, December 15, 1960. The research leading to this paper was supported by the Air Force Missile Dev. Ctr., Holloman Air Force Base, under Contract No. AF 29(600)-1933. J. H. Gengelbach was project supervisor.

† School of Elec. Engrg., Purdue University, Lafayette, Ind.

[1] D. Graham and C. R. Lathrop, "The synthesis of 'optimum' transient response: criteria and standard forms," *Trans. AIEE*, vol. 72 (*Applications and Industry*, no. 2), pp. 273–288; November, 1953.

[2] J. E. Gibson, *et al.*, "Specifications and Data Presentation in Linear Control Systems," Purdue University, School of Elec. Engrg., Lafayette, Ind., Rept. No. 1, AF 29(600)-1933; 1959.

[3] G. C. Newton, L. A. Gould, and J. F. Kaizer, "Analytical Design of Linear Feedback Controls," J. Wiley and Sons, Inc., New York, N. Y., pp. 46–50; 1957.

of gain and time constants of the actual system. It is then a simple matter to calculate the numerical values of the free-gain and time-constant parameters for the optimum system (*i.e.*, to minimize I). The procedure of optimization is illustrated by means of an example of a third-order system.

The General Performance Index

A performance index which enables one to specify the desired response of the optimum system in terms of the differential equation describing the response of an ideal model was proposed by Aizerman.[4] This performance index

$$I = \int_0^\infty \left[e^2 + \tau_1^2 \dot{e}^2 + \tau_2^4 \ddot{e}^2 + \cdots + \tau_n^{2n} \left(\frac{d^n e}{dt^n} \right)^2 \right] dt \quad (2)$$

contains the system-error e, its first n-time derivatives, and the constants τ_i ($i = 1, 2, \cdots, n$) determined from the differential equation of the ideal (=desired) system response. It was shown that the simplest case of the above performance index

$$I = \int_0^\infty [e^2 + \tau_1^2 \dot{e}^2] dt \quad (3)$$

yields optimum systems whose response approaches the response of the first-order model described by the differential equation

$$x + \tau_1^2 \frac{dx}{dt} = 0 \quad (4)$$

as $I \to I_{\min}$. Furthermore, the maximum deviation of the optimum response (when $I = I_{\min}$) from the ideal response (4) is given by the relationship

$$|\Delta x| \leq \sqrt{\frac{I_{\min} - I_{\min\,\min}}{\tau_1^2}}, \quad (5)$$

where Δx represents the maximum difference between the actual response $x(t)$ of the optimum system ($I = I_{\min}$) and the response of an ideal system as shown in Fig. 1. The value of the performance index $I = I_{\min\,\min}$ represents the ideal system obtained by adjusting all the parameters of the system to minimize the performance index I (3).

The importance of (4) is the fact that it enables one to interpret the physical significance of the constant τ_1 in the definition of this particular performance index (3). Thus, while specifying the performance index, one is able to select an ideal system which the response of the actual system will tend toward as the performance index I is minimized.

The ideal model of the first-order system described by (5) is unsatisfactory, however, since it generally does not give enough flexibility to describe the ideal desired response. The performance index

$$I = \int_0^\infty [x^2 + \tau_1^2 \dot{x}^2 + \tau_2^4 \ddot{x}^2] dt \quad (6)$$

yields the ideal model in the form of a second-order system response. To obtain the differential equation describing this second-order model, (6) is rewritten as

$$I = \int_0^\infty [x + \sqrt{\tau_1^2 + 2\tau_2^2}\, \dot{x} + \tau_2^2 \ddot{x}]^2 dt + 2\sqrt{\tau_1^2 + 2\tau_2^2} \int_{x(\infty)}^{x(o)} x dx + 2\tau_2^2 \int_{x\dot{x}(\infty)}^{x\dot{x}(o)} d(x\dot{x}) + 2\tau_2^2 \sqrt{\tau_1^2 + 2\tau_2^2} \int_{\dot{x}(\infty)}^{\dot{x}(o)} \dot{x} d\dot{x}. \quad (7)$$

Evaluating the last three integrals in the above equation, one obtains

$$I = \int_0^\infty [x + \sqrt{\tau_1^2 + 2\tau_2^2}\, \dot{x} + \tau_2^2 \ddot{x}]^2 dt + \sqrt{\tau_1^2 + 2\tau_2^2}\, x^2(o) + 2\tau_2^2 x(o)\dot{x}(o) + \tau_2^2 \sqrt{\tau_1^2 + 2\tau_2^2}\, \dot{x}^2(o). \quad (8)$$

This result is valid, of course, only in asymptotically stable systems, *i.e.*, in systems with

$$x(\infty) = \dot{x}(\infty) = \ddot{x}(\infty) = 0.$$

The requirement of asymptotic stability does not represent any practical limitations, since it is meaningless to

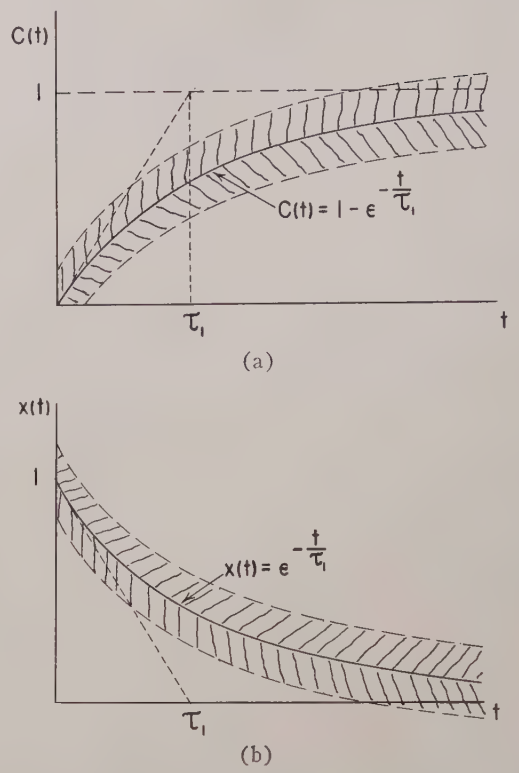


Fig. 1—Ideal response and the region of actual response of Aizerman's optimum system. (a) System error. (b) System response.

[4] M. A. Aizerman, "Lectures on the Theory of Automatic Control," Gostekizdat, Moscow, USSR, 2nd. ed., pp. 302–320; 1958. (In Russian.)

talk about an optimum control system which is either a perfect oscillator or which is unstable.

The ideal system, corresponding to the absolute minimum value $I_{\min\ \min}$ of (8), is described by the equation

$$x + \sqrt{\tau_1^2 + 2\tau_2^2}\,\dot{x} + \tau_2^2\ddot{x} = 0. \tag{9}$$

Comparison of (4) and (9) reveals that the generalization of the performance index of the type of (2) for higher-order models is not obvious, and that each higher-order case would have to be worked out separately. A more serious disadvantage of the performance index (2), proposed by Aizerman, is the limitations imposed on the ideal model. Thus, for example, (9) requires that the ideal model have a damping ratio $\zeta > 0.707$. The same reasoning shows that the performance index of (2) imposes even more severe restrictions on the choice of higher-order ideal models.

In view of the above limitations, one is prompted to look for a more generally applicable performance index. One possibility in this direction would be to define the performance index as

$$I = \int_0^\infty \left[x + \sum_{i=1}^{k} \tau_i \frac{d^i x}{dt^i}\right]^2 dt, \qquad k < n. \tag{10}$$

The absolute minimum of this performance index, and consequently the optimum system, would be obtained when

$$x + \tau_1 \frac{dx}{dt} + \cdots + \tau_k \frac{d^k x}{dt^k} = 0, \qquad k < n. \tag{11}$$

One may observe, however, that (11) represents the characteristic equation (*i.e.*, the denominator of the closed-loop transfer function) of the ideal system. Thus, the above performance index (10) would select as optimum any system which has this particular characteristic equation, regardless of the numerator of the closed-loop transfer function. To differentiate among the systems that have the same characteristic equation and yet differ considerably in their responses due to differences in the numerators of their closed-loop transfer functions, it is necessary that the value of performance index depend upon both the characteristic equation and the initial conditions of the response of the ideal model.

A performance index which meets the above requirements and which is generally applicable with respect to any stable constant-numerator ideal model is proposed in this paper. This performance index is defined as

$$I = \int_0^\infty \left[x^2 + \sum_{i=1}^{k} \tau_i^2 \left(\frac{d^i x}{dt^i}\right)^2 + 2x \sum_{i=2}^{k} \tau_i \frac{d^i x}{dt^i} + 2 \sum_{i=1}^{k} \sum_{j=i+2}^{k} \tau_i \tau_j \frac{d^i x}{dt^i} \frac{d^j x}{dt^j}\right] dt, \qquad k < n \tag{12}$$

where n is the order of the actual system and k is the order of the ideal model.

The error $x(t)$ is defined as the difference between the desired value of the steady-state response c_{ss} and the actual response of the closed-loop system (Fig. 2) $c(t)$; *i.e.*,

$$x(t) = c_{ss} - c(t). \tag{13}$$

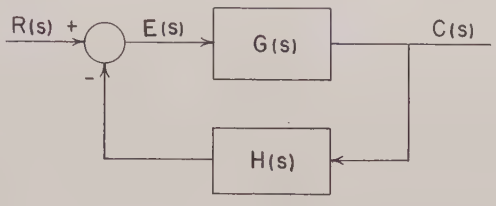


Fig. 2—Block diagram of a linear feedback control system.

Expanding the quadratic term in the integrand of the above equation, one obtains

$$I_k = \int_0^\infty \left[x + \sum_{i=1}^{k} \tau_i \frac{d^i x}{dt^i}\right]^2 dt - 2\int_0^\infty \left[\tau_1 x \frac{dx}{dt} + \sum_{i-1}^{k-1} \tau_i \tau_{i+1} \frac{d^i x}{dt^i} \cdot \frac{d^{i+1} x}{dt^{i+1}}\right] dt, \qquad k < n. \tag{14}$$

If the system is asymptotically stable and if the steady-state error is zero, *i.e.*,

$$x(\infty) = \frac{dx(\infty)}{dt} = \cdots = \frac{d^k x(\infty)}{dt^k} = 0,$$

then (14) is reduced to

$$I_k = \int_0^\infty \left[x + \sum_{i=1}^{k} \tau_i \frac{d^i x}{dt^i}\right]^2 dt + \tau_1 x^2(o) + \sum_{i=1}^{k-1} \tau_i \tau_{i+1} \left[\frac{d^i x(o)}{dt^i}\right]^2, \qquad k < n. \tag{15}$$

The absolute minimum $I_{\min\ \min}$ of this performance index occurs when the integrand of (15) is equal to zero. Consequently, (11) represents the characteristic equation of the ideal model for the proposed performance index (12). The ideal model itself can be represented by a closed-loop system with the transfer function

$$\frac{C(s)}{R(s)} = \frac{1}{\tau_k s^k + \tau_{k-1} s^{k-1} + \cdots + \tau_1 s + 1}. \tag{16}$$

Procedure of System Optimization

The ideal model, with respect to which the system is to be optimized, is described by (11) and (16). Thus, in specifying the performance index for a system, one may first describe the desired response to a unit step or any ether deterministic input in the form

$$c(t) = c_{ss} - \sum_{i=1}^{k} A_i \epsilon - \frac{t}{T_i}, \tag{17}$$

or, according to (13),

$$x(t) = \sum_{i=1}^{k} A_i \epsilon - \frac{l}{T_i} \cdot \tag{18}$$

Substitution of (18) and its first k-time derivatives into (11) yields the constants τ_i of the ideal model. If the desired response (*i.e.*, the response of the ideal model) is specified graphically, one has to find the transfer function (16) of the ideal model. To do this, one has to find the function $c(s)$ whose inverse Laplace transform approximates the specified desired response $c(t)$. Both numerical and graphical techniques for time domain approximations of transfer functions have been reported in the literature.[5]

The next step is to determine the particular performance index I_k (12). In order to accomplish this, let

$$\begin{aligned} I_k &= \lim_{t\to\infty} \int_0^t W\left(x_1 \frac{dx}{dt}, \cdots \frac{d^k x}{dt^k}\right) dt \\ &= \lim_{t\to\infty} V(t) - V(o). \end{aligned} \tag{19}$$

In order to evaluate $V(t)$, one may assume it to be of the type of the quadratic form

$$V = a_{11}x^2 + \sum_{j=2}^{n} a_{1j} x \frac{d^{j-1}x}{dt^{j-1}} + \sum_{i=2}^{n} \sum_{j\geq i}^{n} a_{ij} \frac{d^{i-1}x}{dt^{i-1}} \frac{d^{j-1}x}{dt^{j-1}} \cdot \tag{20}$$

From (19) it is apparent that

$$W = \frac{dV}{dt} \cdot \tag{21}$$

Thus, in order to calculate W, one has to differentiate the function V of (20) with respect to time, and then replace $d^n x/dt^n$ by the lower-order derivatives of x, utilizing the differential equation describing the transient actuating error signal of the actual system, *i.e.*, utilizing the characteristic equation of the actual system. Let the characteristic equation of the actual system be

$$\frac{d^n x}{dt^n} + b_{n-1} \frac{d^{n-1}x}{dt^{n-1}} + \cdots + b_1 \frac{dx}{dt} + b_0 x = 0. \tag{22}$$

This procedure yields the expression for W of the form

$$W = A_{11}x^2 \sum_{j=2}^{n} A_{1j} x \frac{d^{j-1}x}{dt^{j-1}} + \sum_{i=2}^{n} \sum_{j\geq i}^{n} A_{ij} \frac{d^{i-1}x}{dt^{i-1}} \frac{d^{j-1}x}{dt^{j-1}} \cdot \tag{23}$$

Equating the constants A_{ij} of (23) with the corresponding terms of the integrand of (12), one obtains a set of

$$n + (n-1) + (n-2) + \cdots + 1$$

linear algebraic equations containing all a_{ij}'s. Solving these equations for the constants a_{ij} and substituting into (20), one obtains an expression for the function V.

One may note, however, that for asymptotically stable systems

$$x(\infty) = \dot{x}(\infty) = \cdots = x^{(n)}(\infty) = 0, \tag{24}$$

and consequently,

$$\lim_{t\to\infty} V(t) = 0. \tag{25}$$

Hence, the performance index I_k can be expressed as

$$I_k = -V(o). \tag{26}$$

Thus, to calculate the numerical value of I_k, one substitutes the initial values of x and its time derivatives into the expression for the V function.

It is important to note that both the ideal and the actual systems must be stable [*i.e.*, (24) must be satisfied] in order that the results of the optimization procedure be valid. Otherwise, the procedure may yield seemingly reasonable—though incorrect—answers. The procedure used to evaluate the performance index I_k in this paper is identical to the procedure of constructing Liapunov's functions for linear autonomous systems.[6]

Hence, one may use Liapunov's stability theorem and prove that the actual system is stable by proving that the function V (20) is negative definite and W (21) is positive semidefinite. An alternate approach would be to apply the Routh-Hurwitz test to the actual system.

The procedure of optimization will be illustrated by means of examples.

Example 1

Consider the closed-loop system shown in Fig. 2 with the open-loop transfer functions

$$G(s) = \frac{k}{s(1+s)^2}$$

and $H(s) = 1$. Let the ideal-model response be specified as

$$\ddot{x} + 2\dot{x} + x = 0.$$

Comparison of the above equation with (11) yields the following values for the constants τ:

$$\tau_2 = 1$$

and

$$\tau_1 = 2.$$

Hence, from (12) the performance index becomes

$$\begin{aligned} I_2 &= \int_0^\infty [x^2 + \tau_1^2 \dot{x}^2 + \tau_2^2 \ddot{x}^2 + 2\tau_2 \ddot{x}] dt \\ &= \int_0^\infty [x^2 + 4\dot{x}^2 + \ddot{x}^2 + 2x\ddot{x}] dt. \end{aligned}$$

[5] See, *e.g.*, J. E. Storer, "Passive Network Synthesis," McGraw-Hill Book Co., Inc., pp. 303–315; 1959.

[6] I. G. Malkin, "Theory of Stability of Motion," AEC, Oak Ridge, Tenn., Translation No. 3352, pp. 57–61; 1959.

For this example ($n=3$) the V function (20) is

$$V = a_{11}x^2 + a_{12}x\dot{x} + a_{13}x\ddot{x} + a_{22}\dot{x}^2 + a_{23}\dot{x}\ddot{x} + a_{33}\ddot{x}^2.$$

Differentiation of the above equation with respect to time yields

$$\frac{dV}{dt} = W = 2a_{11}x\dot{x} + a_{12}x\ddot{x} + a_{12}\dot{x}^2 + a_{13}x\dddot{x} + a_{13}\dot{x}\ddot{x} + 2a_{22}\dot{x}\ddot{x} + a_{23}\dot{x}\dddot{x} + a_{23}\ddot{x}^2 + 2a_{33}\ddot{x}\dddot{x}.$$

The closed-loop transfer function for this system is

$$\frac{C}{R} = \frac{k}{s(1+s)^2 + k} = \frac{k}{s^3 + 2s^2 + s + k}.$$

Thus, the system characteristic equation is

$$\dddot{x} = 2\ddot{x} + \dot{x} + kx = 0.$$

Solving this equation for $\dddot{x}$ and substituting into the expression for dV/dt, one obtains

$$\frac{dV}{dt} = W = (-ka_{13})x^2 + (2a_{11} - a_{13} - ka_{23})x\dot{x} + (a_{12} - 2a_{13} - 2ka_{33})x\ddot{x} + (a_{12} - a_{23})\dot{x}^2 + (a_{13} + 2a_{22} - 2a_{23} - 2a_{33})\dot{x}\ddot{x} + (a_{23} - 4a_{33})\ddot{x}^2.$$

Comparing this with the integrand of I_2, one may write

$$-ka_{13} = 1; \qquad a_{12} - 2a_{13} - 2ka_{33} = 2;$$
$$a_{12} - a_{23} = 4; \qquad 2a_{11} - a_{13} - ka_{23} = 0;$$
$$a_{23} - 4a_{33} = 1; \qquad a_{13} + 2a_{22} - 2a_{23} - 2a_{33} = 0.$$

Simultaneous solution of the above equations yields

$$a_{11} = \frac{k^3 + 4k^2 + 3k + 2}{2(k^2 - 2k)}; \qquad a_{22} = \frac{k^2 + 6k}{k^2 - 2k};$$
$$a_{12} = \frac{5k^2 - 8k + 8}{k^2 - 2k}; \qquad a_{23} = \frac{k^2 + 4k + 4}{k^2 - 2k};$$
$$a_{13} = \frac{-1}{k} = \frac{2 - k}{k^2 - 2k}; \qquad a_{33} = \frac{1.5k + 1}{k^2 - 2k}.$$

Note that, for this system, the initial conditions due to a unit step input are

$$x(o) = 1$$
$$\dot{x}(o) = 0$$
$$\ddot{x}(o) = 0.$$

Hence, from (19) and (25)

$$I_0 = V(o) = -\frac{k^3 + 4k^2 + 3k + 2}{2(k^2 - 2k)}x^2(o) = \frac{k^3 + 4k^2 + 3k + 2}{4k - 2k^2}.$$

The optimum system will, obviously, correspond to the minimum value of the performance index I_2. This minimum value occurs when $k=0.43$; thus,

$$I_{2\,\min} = 3.04.$$

The numerical value of the performance index for the ideal model is obtained by substituting (11) into (14). This procedure yields

$$I_{\min\,\min} = 2.00.$$

The response of this optimum system is shown in Fig. 3.

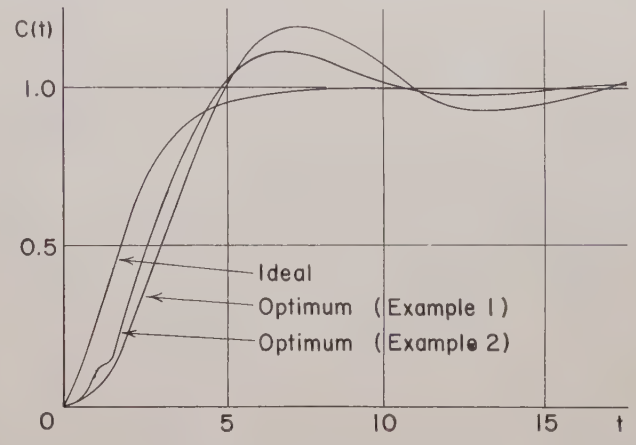


Fig. 3—Ideal response and the responses of optimum systems of Example 1 and Example 2.

Example 2

The preceding example illustrated the procedure of optimization of a system with only one variable parameter. Much better results can be achieved with more free parameters. To illustrate this point, consider the system with

$$G(s) = \frac{0.43}{s(1+s)(1+Ts)}; \qquad 0.5 \gtrless T < \infty,$$

and $H(s)=1$.

By the same procedure used in Example 1, one finds that

$$I_2 = -V(o) = \frac{0.43T^2 + 1.80T + 1.88}{0.49T + 0.86}.$$

With the constraint on T ($0.5 \gtrless T < \infty$), the minimum value of the performance index I_2 occurs at

$$T = 0.5;$$

then

$$I_{2\,\min} = 2.61.$$

The response corresponding to this optimum system is shown in Fig. 3.

The preceding two examples illustrate the procedure of system optimization by means of the proposed performance index. Fig. 3 indicates clearly that the lower the value of the performance index, the closer the optimum system response approaches the specified ideal

response. Without the constraint on the open-loop time-constant T, the second example would yield an unstable "optimum" system with $T<0$. This shows the need to check the stability of the final system, or to constrain the parameter values in such a way as to assure a stable optimum system.

Conclusions

The proposed performance index can be used in a true synthesis procedure of automatic control systems. The importance of the particular performance index proposed in this paper is that it enables one to synthesize an optimum system to external time-domain specifications set up by the user, rather than to some artificial specifications (such as phase margin) chosen by the servo designer.

The examples illustrating the procedure of optimization by means of this performance index, for simplicity, were based upon specifications of system transient response due to step inputs. The procedure of optimization to specified transient response to other deterministic inputs such as ramp, sinusoidal, etc., can be carried out in exactly the same manner; the only difference in such cases would be the change in initial values of response and its time derivatives to be substituted in the expression for the performance index.

A drawback common to the performance index proposed in this paper, and all other integral of error type performance indexes, is that the procedures of evaluating these performance indexes sometimes fail to reveal that the system is unstable and thus yield seemingly reasonable—though erroneous—results. A possibility of incorporating Routh-Hurwitz conditions into the procedure of system optimization is investigated. However, no results in this direction can be reported at the present time.

Acknowledgment

The author is indebted to Dr. J. E. Gibson for his suggestions and critical review of the manuscript.

Stability of Servomechanisms with Friction and Stiction in the Output Element*

P. K. BOHACEK† and F. B. TUTEUR†, member, ire

Summary—**Servomechanisms with friction in the output element are often observed to oscillate, even though the Bode diagram indicates stability. This paper investigates the conditions for this instability and the type of oscillation that can occur. It finds that an overdamped system with a lag equalizer is stable if $L<2C/C-1$, where L is the lag ratio and C = static friction ÷ Coulomb friction; with a lag-lead equalizer it is stable if**

$$\frac{L}{1+a/b}<\frac{2C}{C-1},$$

where a/b is the ratio of the two zeros of the network. For underdamped systems, the same analysis may be carried out, resulting in only slightly more complicated expressions. Experimental results that correlate with the theory are also included.

I. Introduction

THE servomechanism studied in this paper is a single integration, second-order system.[1] All components are linear, except the output power element, which has static friction[2,3] (stiction) and coulomb friction. The block diagram for this system is shown in Fig. 1.

The output element can be represented by the nonlinear differential equation

$$J\frac{d^2\theta}{dt^2}+B\frac{d\theta}{dt}=y-f \tag{1}$$

where J and B are the parameters, such as inertia and viscous friction in a motor. The function f, representing the static and coulomb friction, is a complicated function of the output velocity and the applied force. Its nature is indicated in Fig. 2, as a function of $d\theta/dt$.

The box A represents all the gain of the circuit and is assumed to have a greater bandwidth than the output element.

* Received by the PGAC, December 14, 1960; revised manuscript received, March 14, 1961.

† Elec. Engrg. Dept., Yale University, New Haven, Conn.

[1] J. L. Bower and P. M. Schultheiss, "Introduction to the Design of Servomechanisms," John Wiley and Sons, Inc., New York, N. Y.; 1958.

[2] J. Tou and P. M. Schultheiss, "Static and sliding friction in feedback systems," *J. Appl. Phys.*, vol. 24, pp. 1210–1217; September, 1953.

[3] R. L. Moruzzi and F. B. Tuteur, "Nonlinear servomechanisms of limited dynamic range," *Trans. AIEE*, vol. 79 (*Applications and Industry*, no. 51), pp. 314–320; November, 1960.

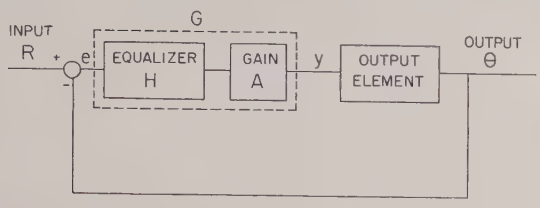


Fig. 1—Block diagram.

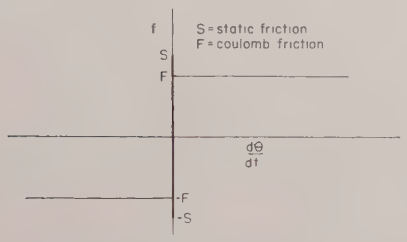


Fig. 2—Static and dynamic friction.

A system of this type often exhibits instability when equalized with a lag or lag-lead equalizer. This instability occurs even though the Bode diagram indicates stability. The Bode diagram is usually measured so as to make friction negligible (for instance at high amplitudes). The oscillations observed are usually of a relaxation type and are of low amplitude and frequency. They are usually at a frequency below the crossover frequency.

A piecewise linear analysis[4] is used to find the condition for oscillation for a lag and a lag-lead equalizer. This is done for the case where the system is underdamped and also where it is overdamped.

By this method it is found that for the under-damped case, stability is achieved if

$$\frac{L}{1+a/b} < \frac{2C}{C-1}, \tag{2}$$

where C = static friction ÷ coulomb friction and a and b are the two zeros of the equalizer. A similar expression is found for the underdamped case.

II. Analysis with Lag Equalizer

A. Overdamped System

Consider a lag network with the transfer function of the form

$$H(s) = \frac{c(s+a)}{a(s+c)}. \tag{3}$$

This network has an assymptotic Bode diagram as sketched in Fig. 3. This type of network has a gain of one for low frequencies and a gain c/a for high frequencies. When a step function of unit amplitude is applied to such a network, the resulting output is sketched in Fig. 4.

[4] W. J. Cunningham, "Introduction to Nonlinear Anlaysis," McGraw-Hill Book Co., Inc., New York, N. Y., pp. 66–69; 1958.

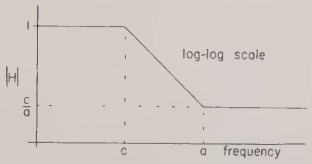


Fig. 3—Bode diagram of lag equalizer.

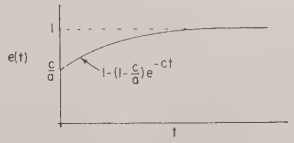


Fig. 4—Step function response of a lag network.

This type of response can be thought of as coming from a network that initially has a gain of c/a and whose gain slowly increases to 1. Applying a step function to this hypothetical network would give the same response as in Fig. 4. The lag equalizer can then be replaced by this type of time varying network.

Let it be assumed that R, the input, changes very slowly. Initially the whole system was at rest and $e=y=\theta=0$. As R changes, so does e and y changes A times faster. The output θ cannot change yet, since y is not large enough to overcome the stiction force, S. Eventually, however, the value of y will overcome the stiction S and the output will start to move. Just before this happens, the following conditions hold:

$$y = S = Ae$$

$$e = \frac{S}{A}$$

$$R = \frac{S}{A}$$

$$\theta = 0.$$

For any rapid motion of θ, the gain of G will be $(c/a)A = A/L$, where L can be identified as the lag ratio. That means that the variation in y, that is, Δy will be

$$\Delta y = -\frac{A}{L}\theta.$$

It has been assumed that R changes so slowly, that during the subsequent motion of θ, it remains constant at $R=S/A$. Then (1) becomes

$$J\frac{d^2\theta}{dt^2} + B\frac{d\theta}{dt} = A\frac{S}{A} + \Delta y - F = S - \frac{A}{L}\theta - F \tag{4}$$

or

$$J\frac{d^2\theta}{dt^2} + B\frac{d\theta}{dt} + \frac{A}{L}\theta = F(C-1) \tag{5}$$

where $C=S/F$.

Laplace transforming both sides gives

$$J\hat{\theta}s^2 + B\hat{\theta}s + \frac{A}{L}\hat{\theta} = \frac{F(C-1)}{s} \tag{6}$$

where F and C are constants, and where $\hat{\theta}=\theta(s)$ is the Laplace transform of θ.

Solving for $\hat{\theta}$

$$\hat{\theta} = \frac{F(C-1)}{s\left(Js^2 + Bs + \frac{A}{L}\right)} \tag{7}$$

$$\hat{\theta} = \frac{LF}{A}(C-1)\left[\frac{1}{s} - \frac{s + \frac{B}{J}}{s^2 + \frac{B}{J}s + \frac{A}{LJ}}\right]. \tag{8}$$

Retransforming into the time domain, $\theta(t)$ becomes

$$\theta(t) = \frac{LF}{A}(C-1)\left[1 - \left(\frac{1}{2} - \frac{B}{2\sqrt{B^2 - 4AJ/L}}\right)\cdot\exp - \left(\frac{B}{2J} + \sqrt{\frac{B^2}{4J^2} - \frac{A}{LJ}}\right)t - \left(\frac{1}{2} + \frac{B}{2\sqrt{B^2 - 4AJ/L}}\right)\cdot\exp - \left(\frac{B}{2J} - \sqrt{\frac{B^2}{4J^2} - \frac{A}{LJ}}\right)t\right]. \tag{9}$$

This equation holds for a short time only, while the gain of G is essentially A. The varying gain of G increases as e^{-ct}, while the response of the system is limited by the time constant of the output element. In most designs the time constant of the equalizer is much greater than the time constant of the system. Hence the system will, to a good approximation, have completed its motion before e^{-ct} has changed much.

Eq. (9) shows that the time response of $\theta(t)$ will be an exponential, with the final value $(LF/A)(C-1)$. If this distance is larger than $2S/A$, then it might be thought that the system would be unstable, since θ would overshoot the desired location by an amount greater than its original deviation S/A.

This condition for instability gives

$$\frac{LF}{A}(C-1) > \frac{2S}{A}. \tag{10}$$

Since $S=CF$, the condition for instability becomes

$$L > \frac{2C}{C-1}. \tag{11}$$

This condition, however, is not quite sufficient for oscillation. Just before θ started to move, y was equal to S and CF. At that time e was S/A. Since θ moved the distance $(LF/A)(C-1)$, then at the end of this motion

$$e = \frac{S}{A} - \theta = \frac{CF}{A} - \frac{LF}{A}(C-1)$$

$$= \frac{F}{A}(C - LC + L). \tag{12}$$

Then y will be

$$y = S - \frac{A}{L}\theta = F.$$

But since eventually y becomes Ae, y will approach the quantity

$$F[L(1-C) + C].$$

When condition (11) is satisfied, this quantity is smaller than $-CF$. Thus at some time y will be equal to $-CF$. Then the conditions will be similar to those just before θ started to move.

The equation of motion will then be similar to (5);

$$J\hat{\theta}s^2 + B\hat{\theta}s + \frac{A}{L}\hat{\theta} = F(1-C). \tag{13}$$

Solving this equation in the identical manner of (5) gives that θ moves exponentially the distance $-(LF/A)(C-1)$. This will bring θ back to its original starting position at $\theta=0$. Then $e=S/A$ and y will approach S. But it will reach this only after an infinite time has passed. Thus it seems that the system will settle down after one cycle.

If it is assumed that for some reason, when θ first moved, it did not move quite far enough, *i.e.*, it moved the distance $(LF/A)(C-1)-i$, where $i>0$, then during the second half of the first cycle, if θ moves $(LF/A)(C-1)$, it will come to rest at $\theta=i$. Then $e=(S/A)+i$ and y will approach $S+Ai$. Thus after some finite time, y will be equal to S and θ will move again. From then on the oscillations will continue, even if θ moves only $(LF/A)(C-1)$ every time. Thus it can be seen that once started properly, the oscillations will continue indefinitely. The amplitude of the oscillations will also be constant.

It can be seen that the two halves of the cycle need not be equally long. In fact it seems probable that they will not be of equal duration.

Since the stiction and friction of most elements is somewhat random, it may be expected that the duration of each half cycle will change. If the stiction should change enough, it may also be possible that the oscillations would stop.

The sketches of the expected waveforms from the above analysis are shown in Fig. 5.

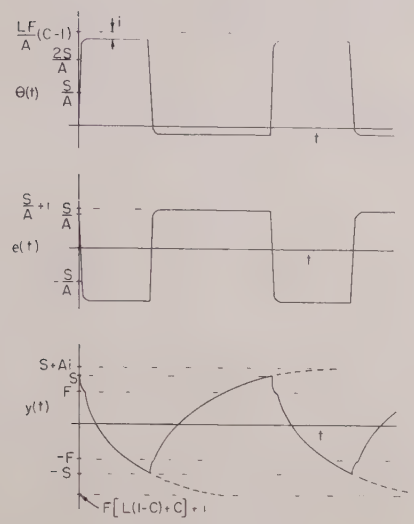


Fig. 5—Waveforms of $\theta(t)$, $e(t)$, and $y(t)$.

B. *Underdamped System*

If the system is underdamped, then in (9)

$$\frac{B^2}{4J^2} < \frac{A}{LJ} \tag{14}$$

and therefore

$$\theta(t) = \frac{LF}{A}(C-1)\left[1 - \sqrt{\frac{A}{LJ}}\,\frac{e^{-Bt/2J}}{\sqrt{\frac{A}{LJ} - \frac{B^2}{4J^2}}} \cdot \cos\left(t\sqrt{\frac{A}{LJ} - \frac{B^2}{4J^2}} - \tan^{-1}\frac{B}{2J\sqrt{\frac{A}{LJ} - \frac{B^2}{4J^2}}}\right)\right]. \tag{15}$$

This solution holds only up to the first maximum, when θ stops for the first time and tries to reverse. When the velocity of θ reverses, the friction reverses. Hence, θ does not start to move again until y is large enough to overcome the stiction, S.

It must then be found where θ stops. The time when this happens can be found by solving $d\theta(t)/dt = 0$ for t. This gives that the time for the first maximum is

$$t_m = \frac{\pi}{\sqrt{\frac{A}{LJ} - \frac{B^2}{4J^2}}}.$$

Substituting this value of $t = t_m$ into (15), $\theta_{\max}$ becomes

$$\theta(t_m) = \theta_{\max} = \frac{LF}{A}(C-1)\left(1 + \exp - \frac{\pi}{\sqrt{\frac{4AJ}{LB^2} - 1}}\right). \tag{16}$$

Then the condition for instability becomes $\theta_{\max} > (2S/A)$. This is equivalent to

$$L\left(1 + \exp - \frac{\pi}{\sqrt{\frac{4AJ}{LB^2} - 1}}\right) > \frac{2C}{C-1}. \tag{17}$$

This condition is the same as in Section II-A, except for the factor

$$\left(1 + \exp - \frac{\pi}{\sqrt{\frac{4AJ}{LB^2} - 1}}\right).$$

This factor varies from 1 to 2, depending on the damping, as long as the system is underdamped, *i.e.*, $(4AJ/LB^2) - 1 > 0$. Eq. (17) reduces to (11) for $(4AJ/LB^2) - 1 = 0$.

III. Analysis with Lag-Lead Equalizer

A. *Overdamped System*

The step function response of a lag-lead equalizer with the transfer function

$$H(s) = \frac{(s+a)(s+b)}{(s+c)(s+d)} \quad \text{where} \quad ab = cd \tag{18}$$

is given by

$$e(t) = 1 + \frac{c + d - b - a}{c - d}(e^{-ct} - e^{-dt}). \tag{19}$$

The waveform of a typical step function response and the assymptotic Bode diagram of a lag-lead network are sketched in Fig. 6.

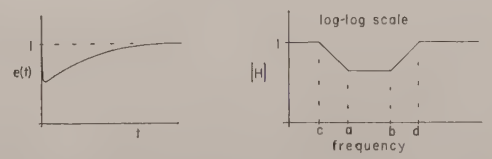


Fig. 6—Step-function response and Bode diagram of lag-lead network.

Comparing this response to that of the lag network, shown in Fig. 4, it may be noticed that they are very similar, except for the sharp spike that the lag-lead response has.

A lag-lead network may also be thought of as a circuit whose gain is initially low and increases with time. In addition the short spike must be considered. This may be treated as a delta function.

Disregarding the delta function for the moment, the initial gain of the hypothetical lag-lead network must be equal to the minimum value that the response reaches after the sharp spike. To find this value, the step function response must be differentiated and set equal to 0.

From this it is found that the minimum occurs at $t = -[1/(d-c)] \log c/d$. Therefore

$$e_{\min} = 1 - \frac{c+d-b-a}{d-c}\left[\left(\frac{c}{d}\right)^{c/d-c} - \left(\frac{c}{d}\right)^{d/d-c}\right].$$

Since $c \ll d$

$$e_{\min} \cong 1 - \left[\left(\frac{c}{d}\right)^{c/d} - \frac{c}{d}\right] + \frac{b+a}{d}\left[\left(\frac{c}{d}\right)^{c/d} - \frac{c}{d}\right].$$

For c/d in the range of 0.001 to 0.01, which is a very good approximation to the values encountered:

$$\left(\frac{c}{d}\right)^{c/d} \cong 1 - 4.7\frac{c}{d}.$$

Then

$$e_{\min} \cong \left(\frac{1}{L} + \frac{a}{d}\right)\left(1 - 5.7\frac{c}{d}\right) + 5.7\frac{c}{d}.$$

where $L = d/b = a/c$.
Since $(c/d) \ll 1$

$$e_{\min} \cong \left(\frac{1}{L} + \frac{a}{d}\right) = \frac{1}{L}\left(1 + \frac{a}{b}\right). \tag{20}$$

The instantaneous gain of the lag network was found to be $1/L$. The instantaneous gain of the lag-lead network is greater by a factor $(1+a/b)$.

This result may be substituted into (9) to get the response with a lag-lead equalizer. The total distance that θ moves is then

$$\frac{LF}{\left(1 + \frac{a}{b}\right)A}(C-1).$$

Instability occurs then if

$$L' = \frac{L}{1 + \frac{a}{b}} > \frac{2C}{C-1}. \tag{21}$$

Now the effect of the delta function must be considered. First of all it must be examined if it is a good assumption to say that the short spike may be thought of as a delta function.

In most applications, the equalizer is designed in such a manner, that the highest pole as $s = -d$ is at a frequency above the major time constants of the system. Since the spike decays with the time constant of $1/d$, it will have decayed before the system has had much time to respond. Thus the assumption that the spike may be handled as a delta function of the same area is a good one.

To find the estimate of the area under the spike, the difference between the lag-lead response and the lag step function response may be taken. This difference Δe is

$$\Delta e = \frac{2c - a - c^2/a}{c-d}e^{-ct} - \frac{c-b-a+d}{c-d}e^{-dt}.$$

Since $c \ll d$ and $a \ll d$,

$$\Delta e \cong \frac{d-b}{d}e^{-dt}.$$

Integrating this gives that the area under the spike is

$$D = \frac{d-b}{d^2} = \frac{1}{d}\left(1 - \frac{1}{L}\right).$$

Since the system is linear, the effect of the delta function may be computed separately. Thus the system equation becomes

$$J\hat{\theta}s^2 + B\hat{\theta}s + \frac{A}{L}\hat{\theta} = D.$$

Solving for θ gives

$$\theta(t) = \frac{D}{\sqrt{B^2 - 4AJ/L'}}\left[\exp -\left(\frac{B}{2J} - \sqrt{\frac{B^2}{4J^2} - \frac{A}{L'J}}\right)t - \exp -\left(\frac{B}{2J} + \sqrt{\frac{B^2}{4J^2} - \frac{A}{L'J}}\right)t\right]. \tag{22}$$

Thus the delta function does not change the final value of θ, and hence it does not affect the stability of the system.

Thus (21) gives the stability criterion for the lag-lead equalizer. Since for the lag equalizer b is infinite, (21) becomes identical to (11), the stability criterion for the lag-network. Therefore (21) holds for both types of equalizers.

Fig. 7 shows the stable and unstable regions for various values of C and L'.

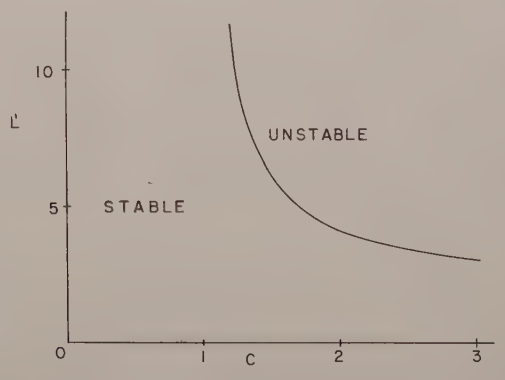


Fig. 7—Stable and unstable regions.

B. Underdamped System

As in the case of the lag network in Section II-B, the lag-lead network case may be analyzed for the underdamped case. In a similar manner the first maximum must be found. Since the same equations hold, the solu-

tion must be the same. Hence the maximum increases by the factor

$$\left(1+\exp-\frac{\pi}{\sqrt{\frac{4AJ}{LB^2}-1}}\right)$$

in the underdamped case. Thus the system is unstable if

$$\frac{L\left(1+\exp-\frac{\pi}{\sqrt{\frac{4AJ}{LB^2}-1}}\right)}{1+\frac{a}{b}}>\frac{2C}{C-1},$$

where $(4AJ/LB^2)-1>0$ from (14).

IV. Experimental Results

A system was built that satisfied the description in Section I. It consisted of a vane-type pneumatic motor, driven by a three-land valve. There was friction and stiction in both the valve and the motor, but that of the motor was much greater.

Lag-lead equalizers with different values of $L/(1+a/b)$ were built. These values were 4, 7, 10 and 14. The value of C was found to be somewhere between 1.4 and 1.6. Hence from the above analysis the system should be unstable for values of L' higher than 6.2 ± 0.8. The experimental results confirmed this prediction. When L' was 10 or 14, the system was definitely unstable and the waveforms observed coincided closely to those of Fig. 5. For $L'=4$, the system was definitely stable. But when L' was 7, the system appeared on the verge of instability. Fig. 8 shows the experimental waveshapes for $L'=7$.

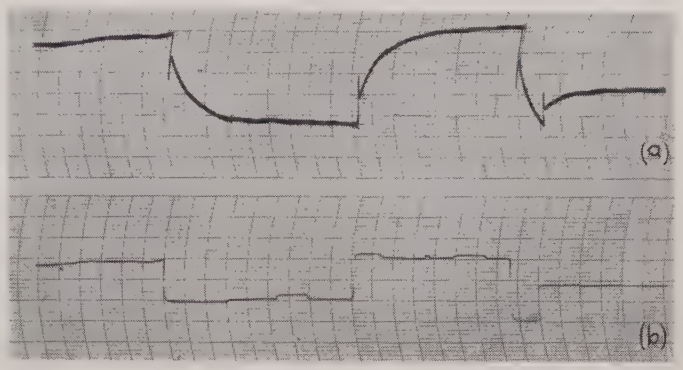


Fig. 8—Experimental results. (a) Waveform of valve position $y(t)$. (b) Waveform of output position $\theta(t)$.

V. Conclusions

In designing a servomechanism where friction and stiction are present, the value of $C=S/F$ must first be found. Then from the expression $2C/C-1$ the largest value of

$$\frac{L\left(1+\exp-\frac{\pi}{\sqrt{\frac{4AJ}{LB^2}-1}}\right)}{1+a/b}$$

can be found. The design of the equalizer must then be made to satisfy the stability condition. Since very often it may be difficult to determine the quantity

$$\exp-\frac{\pi}{\sqrt{\frac{4AJ}{LB^2}-1}},$$

it may be assumed that very little damping is present. In that case the system will be stable if

$$\frac{2L}{1+a/b}>\frac{2C}{C-1}.$$

This will take care of the worst possible case.

Sensitivity Considerations for Time-Varying Sampled-Data Feedback Systems*

J. B. CRUZ, Jr.†, MEMBER, IRE

Summary—A synthesis procedure for linear time-varying sampled-data feedback systems is described. Just as in the continuous system case, one of the advantages of feedback in a sampled-data system is that it can potentially reduce the effect of plant variations on the system performance. To a certain extent, load disturbance and instrument noise may be simultaneously reduced also. A time-domain sensitivity matrix is defined and used in the design of the digital compensators for prescribed insensitivity of the system to plant variation. In addition, two optimization criteria are presented for the design of these compensators when load disturbance and instrument noise have to be reduced as well. The procedure is also applicable to time-invariant sampled-data systems.

Introduction

THE USE of sensitivity functions in the design of linear and continuous feedback systems is well known.[1–3] In this paper, the usefulness of the sensitivity concept is extended to the case where the signals in the feedback system are sampled, and where some or all of the components may vary with time. A mechanical system with some mass which decreases with time is an example of a time-varying system.

Consider the block diagram shown in Fig. 1 where P is the time-varying plant, G is the feed-forward digital compensator,[4] and H is the feedback digital compensator.[4] It is assumed that signal can reach the output only through the plant. Furthermore, there is noise N_1 representing load or output disturbance, and feedback can be obtained only by allowing additional contamination N_2 representing feedback instrumentation noise. This structure has essentially two degrees of freedom,[3] and will be sufficient for illustrating the role of sensitivity in design procedures. The problem is to synthesize the digital compensators G and H given: 1) characteristics of the time-varying plant together with expected variations from the nominal, 2) desired over-all input-output characteristic, 3) noise N_1 and N_2, and 4) an optimization criterion.

The characterization for the various components in the system that is used in this paper is Friedland's transmission matrix.[5] The transmission matrix is essentially the mathematical transformation which relates a sequence of input signal values at the sampling instants to a sequence of output signal values at the sampling instants when the system is linear. The components are assumed to be preceded and followed by sampling switches which are not necessarily ideal impulse modulators.[4] If $x(t)$ is the input to the sampling switch and $x^*(t)$ is the output of the sampling switch, then the transfer characteristic of the switch is assumed to be

$$x^*(t) = \sum_{k=0}^{k=[t/T]} x(kT)s(t - kT), \tag{1}$$

where T is the uniform sampling interval, $[t/T]$ is the largest integer in t/T, and $s(t-kT)$ is the output pulse of the switch when the input is a unit amplitude sample at $t=kT$.

If the response of a physical linear system H, such as that in Fig. 2, due to an input $s(t-\tau)$ is $h(t, \tau)$, then it

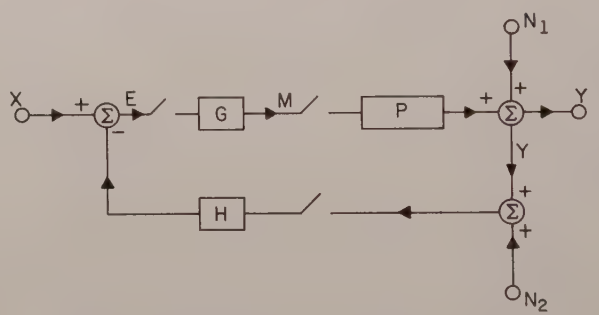


Fig. 1—A digital feedback control system with two degrees of freedom and with load disturbance and feedback instrumentation noise shown.

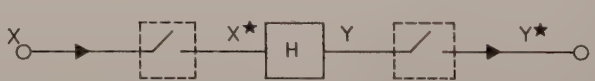


Fig. 2—Block diagram for system with samplers at input and at output.

* Received by the PGAC, December 14, 1960; revised manuscript received, March 13, 1961. This research was supported jointly by the Dept. of the Army (Signal Corps and Ordnance Corps), Dept. of the Navy (Office of Naval Research), and the Dept. of the AF (Office of Scientific Research, Air Res. and Dev. Command) under Signal Corps Contract No. DA-36-039-SC-85122 with the Coordinated Science Lab., University of Illinois, Urbana.

† Elec. Engrg. Dept. and Coordinated Science Lab., University of Illinois, Urbana, Ill.

[1] H. W. Bode, "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Co., Inc., New York, N. Y.; 1945.

[2] J. G. Truxal and I. M. Horowitz, "Sensitivity considerations in active network synthesis," *Proc. 2nd Midwest Symp. on Circuit Theory*, pp. 6-1 to 6-11; December 1956.

[3] I. M. Horowitz, "Fundamental theory of automatic linear feedback control systems," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-4, pp. 5–19; December, 1959.

[4] J. T. Tou, "Digital and Sampled-Data Control Systems," McGraw-Hill Publishing Co., Inc., New York, N. Y.; 1959.

[5] B. Friedland, "A technique for the analysis of time-varying sampled-data systems," *Trans. AIEE, (Applications and Industry)*, vol. 75, pt. 2, pp. 407–414; January, 1957.

follows from the defining property of linear systems that the output $y(t)$ due to an input $x^*(t)$ given by (1) is

$$y(t) = \sum_{k=0}^{k=[t/T]} x(kT)h(t, kT). \tag{2}$$

If only sampling instants are of interest, we may write (2) in matrix form as

$$\begin{bmatrix} y(0) \\ y(T) \\ y(2T) \\ \vdots \end{bmatrix} = \begin{bmatrix} h(0,0) & 0 & \cdots & & \\ h(T,0) & h(T,T) & 0 & \cdots & \\ h(2T,0) & h(2T,T) & h(2T,2T) & \cdots & \\ \vdots & & & & \end{bmatrix} \begin{bmatrix} x(0) \\ x(T) \\ x(2T) \\ \vdots \end{bmatrix}. \tag{3}$$

The matrix

$$H = \begin{bmatrix} h(0,0) & 0 & \cdots & & \\ h(T,0) & h(T,T) & 0 & \cdots & \\ h(2T,0) & h(2T,T) & h(2T,2T) & 0 & \cdots \\ \vdots & & & & \end{bmatrix} \tag{4}$$

is Friedland's transmission matrix of the linear system consisting of the input sampling switch and the system H. From the switch characteristic of (1) the output $y^*(t)$ is related to $y(t)$ by

$$\begin{bmatrix} y^*(0) \\ y^*(T) \\ \vdots \end{bmatrix} = \begin{bmatrix} s(0) & 0 & & & \\ s(T) & s(0) & 0 & \cdots & \\ s(2T) & s(T) & s(0) & 0 & \cdots \\ & & & & \end{bmatrix} \cdot \begin{bmatrix} y(0) \\ y(T) \\ \vdots \end{bmatrix}, \tag{5}$$

which is the matrix analog of (1). From (3) and (5) we have the matrix equation

$$[y^*] = [s][h][x]. \tag{6}$$

All the components will always be preceded by input sampling switches. In many practical cases the pulse $s(t-\tau)$ is very narrow and hold circuits[4] are inserted between the samplers and the following continuous type system. The hold circuit may be considered part of the switch or part of the continuous system.

Fig. 3 (next page) shows the surface $h(t, \tau)$ and the sample points on the surface which represent the elements of the transmission matrix. For $t<\tau$, $h(t, \tau)=0$ because of physical realizability condition. Clearly, for a physically realizable transmission matrix $H=[h_{ij}]$,

$$h_{ij} = 0 \quad \text{for} \quad i < j, \tag{7}$$

so that H must be a triangular matrix. If $h_{ij}=h_{i+k,j+k}$ for all i, j and k which are integers, then we say that the system is time-invariant. Note that from the transmission matrix of the sampler as in (5), the sampler is clearly time-invariant.

Let the plant transmission matrix be denoted by P, that of the feedforward compensator by G, that of the feedback compensator by H, that of the desired over-all system by F_d, and that of the actual over-all system by F. Let Y denote the output signal vector

$$Y = \begin{bmatrix} y(0) \\ y(T) \\ y(2T) \\ \vdots \end{bmatrix}; \tag{8}$$

X the input signal vector,

$$X = \begin{bmatrix} x(0) \\ x(T) \\ x(2T) \\ \vdots \end{bmatrix}; \tag{9}$$

N_1 the load disturbance noise vector,

$$N_1 = \begin{bmatrix} n_1(0) \\ n_1(T) \\ n_1(2T) \\ \vdots \end{bmatrix}; \tag{10}$$

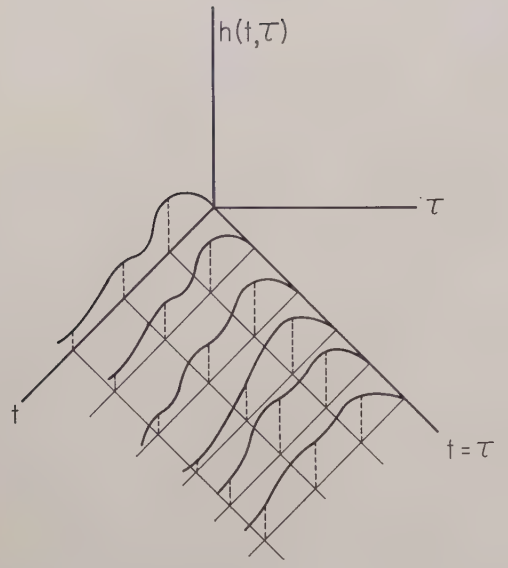


Fig. 3—Impulse response surface $h(t, \tau)$ and sample points representing the elements of the transmission matrix.

and N_2 the feedback instrumentation noise vector

$$N_2 = \begin{bmatrix} n_2(0) \\ n_2(T) \\ n_2(2T) \\ \vdots \end{bmatrix}. \tag{11}$$

From Fig. 1, assuming $N_1 = N_2 = 0$ for the moment, the error vector E is

$$E = X - HY. \tag{12}$$

The plant input vector is

$$M = GE = GX - GHY, \tag{13}$$

and the output Y is

$$Y = PM = PGX - PGHY \tag{14}$$

or

$$Y = (I + PGH)^{-1}PGX, \tag{15}$$

provided $I+PGH$ is nonsingular. Since by definition

$$Y = FX \quad \text{for} \quad N_1 = N_2 = 0, \tag{16}$$

then

$$F = (I + PGH)^{-1}PG. \tag{17}$$

Similarly, assuming $X = N_2 = 0$, and $N_1 \neq 0$,

$$E = - HY, \tag{18}$$

$$M = - GHY, \tag{19}$$

and

$$Y = PM + N_1 = - PGHY + N_1, \tag{20}$$

or

$$Y = (I + PGH)^{-1}N_1, \tag{21}$$

provided $I+PGH$ is nonsingular. The transmission to the output due to N_1 is defined as F_{N_1}, *i.e.*,

$$Y = F_{N_1}N_1 \quad \text{for} \quad X = N_2 = 0, \tag{22}$$

so that

$$F_{N_1} = (I + PGH)^{-1}. \tag{23}$$

For

$$N_1 = X = 0, \qquad N_2 \neq 0,$$

we find

$$E = - H(Y + N_2), \tag{24}$$

$$M = GE = - GHY - GHN_2, \tag{25}$$

and

$$Y = PM = - PGHY - PGHN_2, \tag{26}$$

or

$$Y = - (I + PGH)^{-1} PGHN_2, \tag{27}$$

provided $I+PGH$ is nonsingular. Again the transmission to the output due to N_2 is defined as F_{N2} and from (27),

$$F_{N_2} = - (I + PGH)^{-1} PGH. \tag{28}$$

If X, N_1, and N_2 are all present,

$$Y = FX + F_{N_1}N_1 + F_{N_2}N_2. \tag{29}$$

which is of course a consequence of the superposition property of linear systems.

The Sensitivity Matrix

Let the actual transmission matrices be denoted by the subscript 2, the nominal ones without subscript, ΔP the deviation between P_2 and P, and ΔF the resulting deviation between F_2 and F as caused by the deviation in P, *i.e.*,

$$P_2 = P + \Delta P, \tag{30}$$

$$F_2 = F + \Delta F. \tag{31}$$

ΔP is usually the tolerance on P. The sensitivity matrix will be defined as

$$S = F_2^{-1}\Delta F, \tag{32}$$

where F_2^{-1} is the inverse of F_2 provided F_2 is nonsingular. This sensitivity matrix has the same form as Horowitz's[6] sensitivity matrix for linear time-invariant continuous systems, defined in the frequency domain. Although physically the matrices in Horowitz's paper have an entirely different meaning compared to the

[6] I. M. Horowitz, "Synthesis of linear, multivariable feedback control systems," IRE Trans. on Automatic Control, vol. AC-5, pp. 94–105; June, 1960.

transmission matrices, the algebra is exactly the same. A further generalization of the sensitivity definition in (32) is obtained, if general operators are used instead of transmission matrices.[7]

It can easily be shown that in addition to (17), F is also expressible as[8]

$$F = PG(I + HPG)^{-1}. \tag{33}$$

Similarly,[9]

$$S = F_2^{-1}\Delta F = G^{-1}P_2^{-1}\Delta P P^{-1}F. \tag{34}$$

Replacing ΔP by $P_2 - P$,

$$\begin{aligned} S &= G^{-1}P_2^{-1}(P_2 - P)P^{-1}F \\ &= G^{-1}P_2^{-1}P_2P^{-1}F - G^{-1}P_2^{-1}PP^{-1}F \\ &= G^{-1}(P^{-1} - P_2^{-1})F. \end{aligned} \tag{35}$$

If P, P_2, F, and S are specified, G must be synthesized as

$$G = (P^{-1} - P_2^{-1})FS^{-1}. \tag{36}$$

Similarly, from (17), H must be synthesized as

$$H = F^{-1} - (PG)^{-1}. \tag{37}$$

Eqs. (36) and (37) are the basic design equations for the digital feed-forward and feedback compensators if no noise specifications are involved. That is, for given ΔP and P, if F and S are specified, the synthesis can be done exactly because there are two degrees of freedom. If in addition F_{N_1} and F_{N_2} are specified, in general an exact solution is not possible. Some compromise design must be chosen, and in the next section, we will formulate criteria for optimization. First let us show that the requirements that F_{N_1} and F_{N_2} have arbitrarily small elements are contradictory. From (23),

$$F_{N_1} = F(PG)^{-1}, \tag{38}$$

and from (28),

$$F_{N_2} = -FH = -F_{N_1}PGH. \tag{39}$$

But from (23) again,

$$PGH = F_{N_1}^{-1} - I, \tag{40}$$

so that (39) becomes

$$F_{N_2} = -F_{N_1}(F_{N_1}^{-1} - I) = F_{N_1} - I. \tag{41}$$

Note from (41) that if the elements of F_{N_1} are made much smaller than 1, then F_{N_2} will be essentially a unity transmission matrix. Forcing the elements of F_{N_2} to be very small will result in F_{N_1} being essentially a unit matrix. Thus small F_{N_1} and small F_{N_2} are contradictory. On the other hand, if F_{N_2} is not important, then a design based on forcing S to have small elements is compatible with forcing F_{N_1} to have small elements. From (35) and (38), both the F_{N_1} and S matrices may have as small elements as desired by simply choosing G to have large elements. Thus G can be chosen to have large enough elements so that the resulting elements in S and F_{N_1} are all less than those specified.

Optimization Criteria

In this section, two performance criteria are presented, and designs based on minimizing the associated performance functions are defined as optimum. The complexity of the design procedure and amount of calculation required will usually depend on the choice of a performance function.

Suppose the noise vectors N_1 and N_2 are appreciable and it is desired to reduce their effects at the output. It was shown in the last section that F_{N_1} and F_{N_2} cannot be simultaneously reduced arbitrarily. A suitable compromise performance function is

$$C = K\|F_{N_1}\|^2 + \|F_{N_2}\|^2, \tag{42}$$

where C is the performance function, K is a constant equal to the ratio of noise energy in source N_1 to that in source N_2, $\|F_{N_1}\|^2$ is the square of the Euclidean norm of matrix F_{N_1}, and similarly, $\|F_{N_2}\|^2$ is the square of the Euclidean norm of matrix F_{N_2}. The Euclidean norm of a matrix is defined as

$$\|A\| = \left(\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}^2\right)^{1/2}, \tag{43}$$

which is the square root of the sum of the squares of all its elements. Let the elements of the matrices F_{N_1} and F_{N_2} be denoted by ${}_{N_1}f_{ij}$ and ${}_{N_2}f_{ij}$ respectively where i indicates the row location and j the column location. From (41) and (43), (42) becomes

$$\begin{aligned} C = K\|F_{N_1}\|^2 + \|F_{N_1} - I\|^2 &= K\sum_{i=1}^{N}\sum_{j=1}^{N} {}_{N_1}f_{ij}^2 \\ &+ \left[\sum_{i\neq j}^{N}\sum^{N} {}_{N_1}f_{ij}^2 + \sum_{i=1}^{N}({}_{N_1}f_{ii} - 1)^2\right]. \end{aligned} \tag{44}$$

The problem now is to choose the elements of the F_{N_1} matrix such that C is minimized. Eq. (44) may be further written as

$$\begin{aligned} C = (K+1)\sum_{i\neq j}^{N}\sum^{N} {}_{N_1}f_{ij}^2 \\ + \sum_{i=1}^{N}[K\,{}_{N_1}f_{ii}^2 + ({}_{N_1}f_{ii} - 1)^2]. \end{aligned} \tag{45}$$

It is clear that for $i \neq j$, the optimum choice for ${}_{N_1}f_{ij}$ is zero, *i.e.*,

$${}_{N_1}f_{ij} = 0, \quad \text{for} \quad i \neq j. \tag{46}$$

[7] J. B. Cruz, Jr., "Some techniques for the analysis and synthesis of nonstationary networks," 1961 IRE International Convention Record, pt. 4, to be published.

[8] Horowitz, *op. cit.*, Eq. (8) with T replaced by F.

[9] *Ibid.*, Eq. (13).

Only the main diagonal elements remain to be determined. Differentiating C with respect to f_{jj} and equating to zero,

$$2K_{N_1}f_{jj} + 2({}_{N_1}f_{jj} - 1) = 0 \tag{47}$$

or

$${}_{N_1}f_{jj} = \frac{1}{K+1} \quad \text{for} \quad j = 1, 2, \cdots N. \tag{48}$$

From (38), the feedforward compensator transmission matrix G may now be obtained:

$$G = P^{-1}F_{N_1}{}^{-1}F. \tag{49}$$

Since F_{N_1} is diagonal, its inverse is simply a diagonal matrix whose elements are the reciprocals of those of F_{N_1}. Consequently, from (48) and (49), G becomes

$$G = (K+1)P^{-1}F. \tag{50}$$

Similarly, from (37) and (38),

$$H = F^{-1} - F^{-1}F_{N_1} = F^{-1}[I - F_{N_1}]$$

$$= \frac{K}{K+1}F^{-1}. \tag{51}$$

Eqs. (50) and (51) are the design equations for the digital compensators if the optimization criterion is the minimization of the performance function in (42). The corresponding sensitivity matrix for this design is

$$S = G^{-1}(P^{-1} - P_2{}^{-1})F = \frac{1}{K+1}F^{-1}P(P^{-1} - P_2{}^{-1})F$$

$$= \frac{1}{K+1}F^{-1}[I - P(P+\Delta P)^{-1}]F$$

$$= \frac{1}{K+1}[I - F^{-1}P(P+\Delta P)^{-1}F]. \tag{52}$$

Note that for K large (N_1 much more significant than N_2), (48) and (52) indicate that the elements of F_{N_1} and S are inversely proportional (approximately) to K. This is in agreement with the previous section where the elements of S and F_{N_1} may be made arbitrarily small if N_2 is not important. Furthermore, if $\Delta P \to 0$, then S approaches a null or zero matrix also, which should be as expected.

The second criterion that is presented here is the minimization of the mean square of some error. Let

$$Y_d = F_d X \tag{53}$$

represent the desired output column vector whose elements are the desired outputs at the sampling instants, *i.e.*, $y_d(0)$, $y_d(T)$, $y_d(2T)$, $\cdots$, $y_d(NT-T)$, where F_d is the desired over-all transmission matrix defined previously. Let E represent the difference between the actual output and the desired output, *i.e.*,

$$E = Y - Y_d. \tag{54}$$

Then, if E' is the transpose of E,

$$E'E = (Y - Y_d)'(Y - Y_d) = \sum_{k=0}^{N-1}[y(kT) - y_d(kT)]^2. \tag{55}$$

Let the signal X, the noises N_1 and N_2, and the plant perturbation ΔP have random elements (ΔP here has a different meaning from that in the previous section). The random entries in ΔP may either represent errors in measurement or actual random fluctuations in the plant P. In order to have a meaningful performance function, some statistical operation must be performed on $E'E$. In particular, the performance function will be taken as the expected value of $E'E$ with respect to X, N_1, N_2, and ΔP, *i.e.*,

$$C = \langle\langle\langle\langle EE\rangle_X\rangle_{N_1}\rangle_{N_2}\rangle_{\Delta P}. \tag{56}$$

The performance function in (56) is the same function which Fleischer[10] applied to the continuous, linear, time-invariant case. The optimization problem is now clear: to choose G and H such that C in (56) is minimum. From (29) and (53), (55) becomes

$$E'E = [F_2X + F_{N_1}N_1 + F_{N_2}N_2 - F_dX]'$$
$$\cdot[F_2X + F_{N_1}N_1 + F_{N_2}N_2 - F_dX] \tag{57}$$

or

$$E'E = X'(F_2 - F_d)'(F_2 - F_d)X + N_1'F_{N_1}'F_{N_1}N_1$$
$$+ N_2'F_{N_2}'F_{N_2}N_2 + X'(F_2 - F_d)'(F_{N_1}N_1 + F_{N_2}N_2)$$
$$+ (N_1'F_{N_1}' + N_2'F_{N_2}')(F_2 - F_d)X + N_1'F_{N_1}'F_{N_2}N_2$$
$$+ N_2'F_{N_2}'F_{N_1}N_1, \tag{58}$$

where primes denote the transpose operation. It is assumed that the signal, the noises, and the plant perturbation are statistically independent so that the order of operation of taking expected values in (56) is of no consequence. It is further assumed that the elements of N_1, N_2, and ΔP have zero mean. Let us investigate the individual terms in (58). For the first one, we have

$$X'(F_2 - F_d)'(F_2 - F_d)X$$
$$= \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} a_{i+1,j+1}x(iT)x(jT), \tag{59}$$

where the $a_{i+1,j+1}$'s are the elements of the symmetric matrix $(F_2 - F_d)'(F_2 - F_d)$. For convenience let us change indices in (59):

$$X'(F_2 - F_d)'(F_2 - F_d)X$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}x(iT - T)x(jT - T). \tag{60}$$

[10] P. E. Fleischer, "Optimum Design of Passive-Adaptive Linear Feedback Systems with Varying Plants," presented at the Joint Automatic Control Conference, M.I.T., Cambridge, Mass.; September 7–9, 1960.

For the second term we have

$$N_1'F_{N_1}'F_{N_1}N_1 = \sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij}n_1(iT-T)n_1(jT-T), \quad (61)$$

where the b_{ij}'s are the elements of the symmetric matrix $F_{N_1}'F_{N_1}$. Similarly,

$$N_2'F_{N_2}'F_{N_2}N_2 = \sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij}n_2(iT-T)n_2(jT-T),$$

$$X'(F_2-F_d)'(F_{N_1}N_1+F_{N_2}N_2) + (N_1'F_{N_1}'+N_2'F_{N_2}')(F_2-F_d)X = \sum_{i=1}^{N}\sum_{j=1}^{N} x(iT-T)[d_{ij}n_1(jT-T)+k_{ij}n_2(jT-T)], \quad (63)$$

and

$$N_1'F_{N_1}'F_{N_2}N_2 + N_2'F_{N_2}'F_{N_1}N_1 = \sum_{i=1}^{N}\sum_{j=1}^{N} l_{ij}n_1(iT-T)n_2(jT-T), \quad (64)$$

where the elements c_{ij}, d_{ij}, k_{ij}, and l_{ij} have obvious meanings. Taking the expected value of $E'E$ in (58) with respect to the x's, n_1's, and n_2's,

$$\langle\langle\langle E'E\rangle_X\rangle_{N_1}\rangle_{N_2} = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\langle x(iT-T)x(jT-T)\rangle + \sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij}\langle n_1(iT-T)n_1(jT-T)\rangle + \sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij}\langle n_2(iT-T)n_2(jT-T)\rangle, \quad (65)$$

where the rest of the terms drop out because the x's, n_1's, and n_2's are assumed to be independent and the n's have zero means. Let the statistical autocorrelation functions of x, n_1, and n_2 be represented respectively by

$$R_x(i,j) = \langle x(iT-T)x(jT-T)\rangle, \quad (66)$$

$$R_{n_1}(i,j) = \langle n_1(iT-T)n_1(jT-T)\rangle, \quad (67)$$

$$R_{n_2}(i,j) = \langle n_2(iT-T)n_2(jT-T)\rangle. \quad (68)$$

Presently, all these functions are supposed to be given. Later on, it turns out that $R_x(i,j)$ is not needed. The final step in determining the performance function is the statistical averaging of (65) with respect to the random fluctuations in ΔP:

$$C = \sum_{i=1}^{N}\sum_{j=1}^{N}[\langle a_{ij}\rangle_{\Delta P}R_x(i,j) + \langle b_{ij}\rangle_{\Delta P}R_{n_1}(i,j) + \langle c_{ij}\rangle_{\Delta P}R_{n_2}(i,j)]. \quad (69)$$

Note that a_{ij}, b_{ij}, and c_{ij} depend on the elements of ΔP which are assumed to be random. In particular, a_{ij}, an element of the matrix $(F_2-F_d)'(F_2-F_d)$, is given by

$$a_{ij} = \sum_{k=\max(i,j)}^{N} ({}_2f_{ki}-{}_df_{ki})({}_2f_{kj}-{}_df_{kj}), \quad (70)$$

where ${}_2f_{ki}$ and ${}_df_{ki}$ are elements of F_2 and F_d, respectively. Similarly,

$$b_{ij} = \sum_{k=\max(i,j)}^{N} ({}_{N_1}f_{ki})({}_{N_1}f_{kj}) \quad (71)$$

and

$$c_{ij} = \sum_{k=\max(i,j)}^{N} ({}_{N_2}f_{ki})({}_{N_2}f_{kj}). \quad (72)$$

The reason for starting the series from $k=\max(i,j)$ is that F, F_{N_1}, and F_{N_2} are triangular matrices and the neglected terms are zero. Before the function C may be minimized with respect to the elements of G and H, the statistical quantities $\langle a_{ij}\rangle$, $\langle b_{ij}\rangle$, and $\langle c_{ij}\rangle$ must be expressed in terms of the nominal values of g_{ij} and h_{ij}. In general, even for simple probability distributions of the elements in ΔP, the resulting expressions are quite complicated. For this reason, approximations will be used. From (33) F_2 is

$$F_2 = F+\Delta F = (P+\Delta P)G[I+H(P+\Delta P)G]^{-1}. \quad (73)$$

The perturbation ΔP will be neglected in the bracket so that

$$F_2 = F+\Delta F \approx (P+\Delta P)G[I+HPG]^{-1}. \quad (74)$$

This is equivalent to linearizing the dependence of ΔF on ΔP. Using (33), (74) may be written as

$$F_2 = (P+\Delta P)P^{-1}F = (I+\Delta PP^{-1})F = F+\Delta PP^{-1}F. \quad (75)$$

This is equivalent to

$${}_2f_{ij} \approx f_{ij} + \sum_{k=1}^{N}\sum_{r=1}^{N} \delta_{ik}P_{kr}f_{rj}, \quad (76)$$

where δ_{ik} is an element of ΔP and P_{kr} is an element of P^{-1}. Hence,

$$a_{ij} = \sum_{k=\max(i,j)}^{N}\left[f_{ki}-{}_df_{ki}+\sum_{r=s}^{N}\sum_{s=i}^{N}\delta_{kr}P_{rs}f_{si}\right] \times\left[f_{kj}-{}_df_{kj}+\sum_{n=m}^{N}\sum_{m=j}^{N}\delta_{kn}P_{nm}f_{mj}\right]. \quad (77)$$

Assuming that the random plant perturbations δ_{pq} have zero mean, and assuming that δ_{kr} is independent of δ_{kn} for $r\neq n$, then $\langle a_{ij}\rangle$ simplifies to

$$\langle a_{ij}\rangle = \sum_{k=\max(i,j)}^{N}\left[(f_{ki}-{}_df_{ki})(f_{kj}-{}_df_{kj}) + \sum_{m=i}^{N}\sum_{s=i}^{N}\sum_{r=\max(s,m)}^{N}\langle\delta_{kr}^2\rangle P_{rs}P_{rm}f_{si}f_{mj}\right]. \quad (78)$$

Similarly, for $\langle b_{ij}\rangle$, an approximate form for the perturbed F_{N_1} will be used. Thus instead of

$${}_2F_{N_1} = (F+\Delta F)[(P+\Delta P)G]^{-1} \quad (79)$$

from (38), we will neglect ΔP in the bracket and use only a linear approximation,

$$ {}_2F_{N_1} \approx (F + \Delta F)(PG)^{-1} \approx (I + \Delta PP^{-1})F(PG)^{-1} \approx (I + \Delta PP^{-1})F_{N_1}. \quad (80) $$

The pre-subscript 2 denotes the perturbed matrix. Denoting the elements of ${}_2F_{N_1}$ by ${}_{N_1}{}^2f_{ij}$ and those of F_{N_1} by ${}_{N_1}f_{ij}$, we have

$$ {}_{N_2}{}^2f_{ij} = {}_{N_1}f_{ij} + \sum_{r=s}^{N}\sum_{s=j}^{N} \delta_{ir}P_{rs}\,{}_{N_1}f_{sj}, \quad (81) $$

so that (71) may now be approximated as

$$ b_{ij} = \sum_{k=\max\,(i,j)}^{N} \left({}_{N_1}f_{ki} + \sum_{r=s}^{N}\sum_{s=i}^{N} \delta_{kr}P_{rs}\,{}_{N_1}f_{si}\right) \cdot \left({}_{N_1}f_{kj} + \sum_{n=m}^{N}\sum_{m=j}^{N} \delta_{kn}P_{nm}\,{}_{N_1}f_{mj}\right) \quad (82) $$

and

$$ \langle b_{ij}\rangle = \sum_{k=\max\,(i,j)}^{N} \left({}_{N_1}f_{ki}\,{}_{N_1}f_{kj} + \sum_{r=\max\,(i,j)}^{N} \sum_{s=i}^{N}\sum_{m=1}^{N} \langle \delta_{kr}{}^2\rangle P_{rs}\,P_{rm}\,{}_{N_1}f_{si}\,{}_{N_1}f_{mj}\right). \quad (83) $$

Using the same approximation as above for F_{N_2} (41) yields

$$ {}_2F_{N_2} \approx (I + \Delta PP^{-1})F_{N_1} - I, \quad (84) $$

so that

$$ {}_{N_2}{}^2f_{ij} = {}_{N_2}f_{ij} + \sum_{r=s}^{N}\sum_{s=j}^{N} \delta_{ir}P_{rs}\,{}_{N_1}f_{sj} - 1 \quad (85) $$

and $\langle c_{ij}\rangle$ is approximately

$$ \langle c_{ij}\rangle = \sum_{k=\max\,(i,j)}^{N} \left[({}_{N_1}f_{ki} - 1)({}_{N_1}f_{kj} - 1) + \sum_{r=s}^{N}\sum_{s=i}^{N}\sum_{m=j}^{N} \langle \delta_{kr}{}^2\rangle P_{rs}P_{rm}\,{}_{N_1}f_{si}\,{}_{N_1}f_{mj}\right]. \quad (86) $$

The approximations for $\langle a_{ij}\rangle$, $\langle b_{ij}\rangle$, and $\langle c_{ij}\rangle$ as given in (78), (83), and (86), respectively, are used for determining the approximate form for C in (69).

It should be noted that the performance function is a function of the unknown f_{ij}'s and ${}_{N_1}f_{ij}$'s, that is, the elements of the nominal F and F_{N_1} matrices. The other quantities in the expression for C are assumed to be known. It is clear that once the matrices F and F_{N_1} are determined, G and H may be subsequently determined from (37) and (38). Let us proceed then to the minimization of C with respect to the f_{ij}'s and ${}_{N_1}f_{ij}$'s. Taking the partial derivative of C with respect to f_{pq} and noting that $R_x(i, j) = R_x(j, i)$,

$$ 2\sum_{j=1}^{p}(f_{pj} - {}_df_{pj})R_x(q,j) + 2\sum_{k=\max\,(j,q)}^{N}\sum_{r=\max\,(m,p)}^{N}\sum_{m=j}^{N}\sum_{j=1}^{N} \langle \delta_{kr}{}^2\rangle P_{rp}P_{rm}f_{mj}R_x(q,j) = 0 \quad \text{for all } p = 1, 2, \cdots N,\ q = 1, 2, \cdots N,\ p \geq q \quad (87) $$

or

$$ \sum_{j=1}^{p} R_x(q,j)\left[f_{pj} + \sum_{k=\max\,(q,j)}^{N}\sum_{r=\max\,(m,p)}^{N}\sum_{m=j}^{N} \cdot \langle \delta_{kr}{}^2\rangle P_{rp}P_{rm}f_{mj} - {}_df_{pj}\right] = 0. \quad (88) $$

It is desirable at this point to obtain a solution which is independent of input signal statistics. From (88), it is clear that if the $R_x(q, j)$'s are arbitrary constants, then the validity of (88) is guaranteed if we set the coefficients of $R_x(q, j)$ to zero for every j. Thus

$$ f_{pj} + \sum_{k=\max\,(r,j)}^{N}\sum_{r=\max\,(m,p)}^{N}\sum_{m=j}^{r} \langle \delta_{kr}{}^2\rangle P_{rp}P_{rm}f_{mj} = {}_df_{pj} \quad (89) $$

for

$$ p = 1, 2, \cdots N; \quad j = 1, 2, \cdots N; \quad p \geq j. $$

Note that k now starts at max (r, j). The reason for this is that δ_{kr} is zero for $k > r$. Hence we only need sum on k in (89) from $k = \max\,(q, j, r)$. Since r starts from $r = \max\,(m, p)$ and since $p \geq q$, then max $(q, j, r) = \max\,(j, r)$. These $N(N+1)/2$ linear equations yield the $N(N+1)/2$ nonzero elements of F. Fortunately, these $N(N+1)/2$ equations are in the special form: one equation in f_{NN}, two equations in $f_{N,N-1}, f_{N-1,N-1}$, three equations in $f_{N,N-2}, f_{N-1,N-2}, f_{N-2,N-2}, \cdots$ and N equations in $f_{N,1}, f_{N-1,1}, f_{N-2,1}, \cdots, f_{11}$. For example, setting $p = N$, and $j = N$ in (89), yields

$$ f_{NN} + \langle \delta_{NN}{}^2\rangle P_{NN}{}^2f_{NN} = {}_df_{NN} \quad (90) $$

or

$$ f_{NN} = \frac{{}_df_{NN}}{1 + \langle \delta_{NN}{}^2\rangle P_{NN}{}^2}. \quad (91) $$

For $p = N$, $j = N-1$,

$$ f_{N,N-1} + \langle \delta_{NN}{}^2\rangle P_{NN}(P_{N,N-1}f_{N-1,N-1} + P_{NN}f_{N,N-1}) = {}_df_{N,N-1} \quad (92) $$

and for $p = N-1$, $j = N-1$,

$$ f_{N-1,N-1} + [\langle \delta_{N-1,N-1}^2\rangle + \langle \delta_{N,N-1}^2\rangle]P_{N-1,N-1}^2 f_{N-1,N-1} + \langle \delta_{NN}^2\rangle P_{N,N-1}^2 f_{N-1,N-1} + \langle \delta_{NN}^2\rangle P_{N,N-1}P_{NN}f_{N,N-1} = {}_df_{N-1,N-1}. \quad (93) $$

The solutions for (92) and (93) are

$$f_{N,N-1} = \{ {}_{d}f_{N,N-1}[1 + (\langle \delta^2_{N-1,N-1} \rangle + \langle \delta^2_{N,N-1} \rangle) P^2_{N-1,N-1} + \langle \delta^2_{NN} \rangle P^2_{N,N-1}] - {}_{d}f_{N-1,N-1} \langle \delta^2_{NN} \rangle P_{NN} P_{N,N-1} \} \cdot \{ (1 + \langle \delta^2_{NN} \rangle P^2_{N,N})[1 + (\langle \delta^2_{N-1,N-1} \rangle + \langle \delta^2_{N,N-1} \rangle) P^2_{N-1,N-1}] + \langle \delta^2_{NN} \rangle P^2_{N,N-1} \}^{-1} \quad (94)$$

and

$$f_{N-1,N-1} = \{ {}_{d}f_{N-1,N-1}(1 + \langle \delta^2_{NN} \rangle P^2_{NN}) - {}_{d}f_{N,N-1} \langle \delta^2_{NN} \rangle P_{N,N-1} P_{NN} \} \{ (1 + \langle \delta^2_{NN} \rangle P^2_{NN}) \cdot [1 + (\langle \delta^2_{N-1,N-1} \rangle + \langle \delta^2_{N,N-1} \rangle) P^2_{N-1,N-1}] + \langle \delta^2_{NN} \rangle P^2_{N,N-1} \}^{-1}. \quad (95)$$

Continuing in this manner the rest of the elements of F are determined.

Similarly, taking the partial derivative of the performance function C with respect to ${}_{N_1}f_{pq}$ results in

$$2 \sum_{j=1}^{p} {}_{N_1}f_{pj} R_{n_1}(q,j) + 2 \sum \sum \sum \sum \langle \delta_{kr}{}^2 \rangle P_{rp} P_{rm\,N_1} f_{mj} R_{n_1}(q,j) + 2 \sum \sum \sum \sum \langle \delta_{kr}{}^2 \rangle P_{rp} P_{rm\,N_1} f_{mj} R_{n_2}(q,j) + 2 \sum ({}_{N_1}f_{pj} - 1) R_{n_2}(q,j) = 0 \quad (96)$$

or

$$\sum_{j=1}^{p} [R_{n_1}(q,j) + R_{n_2}(q,j)][{}_{N_1}f_{pj} + \sum_{k} \sum_{r} \sum_{m} \langle \delta_{kr}{}^2 \rangle P_{rp} P_{rm\,N_1} f_{mj}] = \sum_{j=1}^{p} R_{n_2}(q,j) \quad (97)$$

for $p=1, 2, \cdots N$; $q=1, 2, \cdots N$; $p \geq q$.

The $N(N+1)/2$ unknown elements of F_{N_1} are determined from the $N(N+1)/2$ linear equations in (97). If it is assumed that noise occurring at different instants of time are independent or uncorrelated, then $R_{n_1}(q, j)$ and $R_{n_2}(q, j)$ are zero unless $q=j$. Hence (97) reduces to

$${}_{N1}f_{pq} + \sum_{k=r} \sum_{r=\max\,(m,p)} \sum_{m=q} \langle \delta_{kr}{}^2 \rangle P_{rp} P_{rm\,N_1} f_{mq} = \frac{R_{n_2}(q,q)}{R_{n_1}(q,q) + R_{n_2}(q,q)}, \quad (98)$$

and the problem is similar to that in (89). That is, the $N(N+1)/2$ equations separate into one equation in one unknown, two equations in two unknowns, and so on.

With the F and F_{N_1} matrices determined, (38) yields

$$G = P^{-1} F_{N_1}{}^{-1} F \quad (99)$$

and from (38) and (39),

$$H = F^{-1} - F^{-1} F_{N_1} = F^{-1}[I - F_{N_1}]. \quad (100)$$

Eqs. (99) and (100) are the final design equations. Note that these are the same design equations as obtained in (49) and (51) for the first performance function considered. The difference is that F_{N_1} and F are chosen differently.

In the two performance criteria discussed above, elements of the finite dimensional transmission matrices are involved. The instants of time corresponding to these elements are assumed to be in the future and in the present. Thus for the second performance function as given in (55) and (56), the mean of the sum of the squares of the *predicted* errors is involved. The instant $t=0$ corresponds to the present, T corresponds to the instant T units of time later compared to the present, and so on up to $(N-1)T$ units of time of prediction. The compensators are designed only to minimize errors at the present and in the future, not the past.

Ideally, it is required that new data on P, F_d, ΔP, N_1, and N_2 will become available for subsequent ranges of prediction time. Thus it is envisioned that a sequence of matrices for G and H will be determined so that the digital compensators will be automatically resynthesized every NT time units. The synthesis for a specific range of NT time units is on a short time basis which considers predicted errors in the P matrix by means of the ΔP specification, and also includes effects of noise disturbances. The new batches of data for succeeding ranges of predicting time in a way provide information on long-term changes in the nature of the plant, input signal, and noise.

Realization of G and H

The transmission matrices G and H for the digital compensators may be realized by means of a digital computer. A model which is useful at least for visualization is a tapped delay line with time-varying gain amplifiers at the taps and an adder at the output.[11–13] Another model, either for visualization or actual realization, is shown in Fig. 4. Both the tapped delay line and the model in Fig. 4 have finite memory. That is,

$$h(iT, jT) = 0 \quad \text{for} \quad i - j > M, \quad (101)$$

where MT is the memory of the system. For many physical systems, the elements of the transmission matrix far away from the main diagonal are negligibly small so that (101) is a reasonable approximation.

In Fig. 4, T is the sampling interval which is the dwell plus travel time for the rotating switches between adja-

[11] T. Kailath, "Sampling Models for Linear Time-Variant Filters," Research Lab. of Electronics, M.I.T., Cambridge, Mass., Tech. Rept. No. 352; May 25, 1959.

[12] J. B. Cruz, Jr., "A generalization of the impulse train approximation for time-varying linear system synthesis in the time domain," IRE Trans. on Circuit Theory, vol. CT-6, pp. 393–394; December, 1959.

[13] J. B. Cruz, Jr., and M. E. Van Valkenburg, "The synthesis of models for time-varying linear systems," *Proc. Symp. on Active Networks and Feedback Systems*, Polytech. Inst. Brooklyn, Polytechnic Press, New York, N. Y., pp. 527–544; 1960.

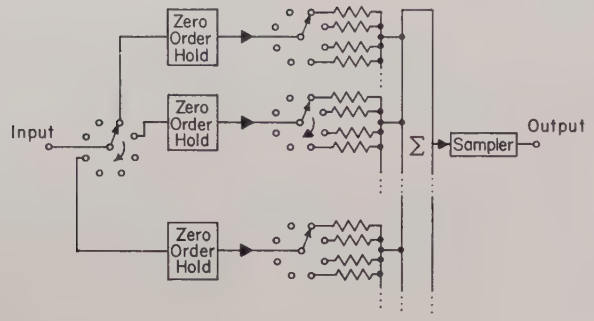


Fig. 4—Model for finite memory digital compensator

cent positions and M is the number of switch positions or contacts in each rotating switch. The zero-order hold circuits have a hold time of MT time units. The input resistors to the adder provide the various scale factors for the different elements $h(iT, jT)$ or $g(iT, jT)$ of the transmission matrices. At the end of MT time units, the resistor values are reset to new ones appropriate for the succeeding interval of MT units of time. The input-output transfer characteristic of the model in Fig. 4 is

$$y(nT) = \sum_{k=n-m}^{n} h(nT, kT)x(kT). \tag{102}$$

Although the model in Fig. 4 uses resistors for the appropriate weights for the corresponding elements in the transmission matrix, the actual hardware will depend on the specific application. For large scale and extensive plants or processes, the digital computer may be the only feasible way. Here, storage or memory takes the place of the hold circuit.

Time-Invariant Transmission Matrices

The techniques described in this paper are valid whether the linear plant is time-varying or not. The method here differs from conventional ones using Z transforms in that the design is carried out entirely in the time domain directly, whereas, if Z transforms are used, specifications will usually have to be converted to the Z domain or some other "frequency" domain first. Note that for the truly time-invariant case, the optimization criteria discussed in the previous section relate to the initial transient. This is because only the first N sampling instants are involved.

Another possible application is in the design of slowly time-varying adaptive systems. If the system variation is slow enough, the P, ΔP, F_d, and F matrices may be considered matrices for time-invariant systems. However, the predicted set of matrices for the next time duration of N samples may be different, so that the effect is a piece-wise or step-wise time-variation. This is active process adaptation.[14]

Conclusion

The advantages of feedback in continuous control systems is well exploited. This does not seem to be the case for sampled-data systems. In this paper some methods of synthesizing the digital compensators in a time-varying system with two degrees of freedom have been presented. These compensators are designed not only to yield a desirable over-all input-output characteristic but also to minimize the effects of plant variations, load disturbance, and instrument noise. The techniques are essentially the same for more degrees of freedom.

The minimization for the second performance function considered in the paper involved some approximation. Consequently, the results are accurate only when the approximations are good enough. In particular, we assumed that the perturbation from the nominal predicted plant transmission matrix is not too large. If this assumption is not justified, then higher-order terms must be included, resulting in nonlinear equations for the determination of the elements of the F and F_{N_1} matrices.

In the application to adaptive control, it is assumed that P is known or can be determined. That is, it is assumed that the so-called identification problem is solved. However, for time-varying plants, especially if the variation is not slow, it is extremely important that P must be for future values of time for the design to be meaningful.

Acknowledgment

The author would like to thank Dr. M. E. Van Valkenburg, C. Pottle, and other members of the Adaptive Systems Group at the Coordinated Science Laboratory, University of Illinois, Urbana, for helpful discussions.

[14] J. A. Aseltine, A. R. Mancini, and C. W. Sarture, "A survey of adaptive control systems," IRE Trans. on Automatic Control, vol. PGAC-6, pp. 102–108; December, 1958.

Direct Cycle Nuclear Power Plant Stability Analysis*

D. BUDEN† MEMBER, IRE, AND R. F. MILLERT†

Summary—A power plant with a heat exchanger such as a nuclear reactor substituted for the conventional chemical interburners in a jet engine will cause a considerable change in dynamic performance. The instantaneous power generated by the heat source is not the same as the instantaneous power delivered to the turbine. The basic control problems are analyzed using fixed control parameters and partial derivatives around a given operating point. A mathematical criterion is developed and correlated with power plant test data.

An understanding of the inherent limitations of combining a reactor, or any heat exchanger having a thermal lag, with a basic jet engine makes it possible to devise a means of control. The introduction of an effective operational speed control makes it possible to operate a complete power plant under any desired condition.

Nomenclature

Symbol

A = Heat transfer area (inch^2)
A_8 = Jet nozzle area (inch^2)
c_{pa} = Specific heat of air, constant pressure (BTU/pound-°R)
c_{pf} = Specific heat of fuel elements, constant pressure (BTU/pound-°R)
h = Film coefficient of heat transfer, fuel elements (BTU/inch^2-sec-°R)
M_f = Mass of fuel elements (pound)
N = Speed (rpm)
Q_a = Heat energy delivered to the air (Mw)
Q_g = Heat energy generated in the reactor (Mw)
T_a = Bulk air temperature (°R)
T_f = Fuel element temperature (°R)
T_3 = Reactor inlet air temperature (°R)
T_4 = Reactor exit air temperature (°R)
t = Time (seconds)
W_a = Reactor airflow (pound/second)
W_{fr} = Afterburner fuel flow (pound/hour)
β = Compressor stator blades
τ = Net torque (foot-pound)
δT = Air temperature change across the reactor (°R)

Note: Subscript zero denotes partial derivative evaluated at a point.

I. Introduction

REPLACING the chemical interburners in a jet engine with a heat exchanger such as a nuclear reactor will cause a considerable change in the dynamic performance. No longer is the instantaneous power generated by the heat source the same as the instantaneous power delivered to the turbine. The basic control problems can be analyzed assuming fixed control parameters and using partial derivatives at a given operating point. The results are correlated with test results by the General Electric Company during a heat transfer reactor experiment using modified J-47-GE-11 turbojet engines and a nonlinear analog power plant simulator.

The analysis is used to examine various methods that can be used to control nuclear power plants. A proper selection of the control parameters will make possible operation at all desired engine speeds.

II. Reactor-Engine Stability Analysis

A basic jet engine (assuming fixed control parameters) consists of a compressor section, burner section, turbine section, tailpipe and nozzle. The compressor is used to compress air drawn through the inlet to a high pressure level. In the burner section of a conventional jet engine, fuel is sprayed into the air and burned. The gas then passes through the turbine which is directly connected to the compressor. From here, the gas is ducted through the tailpipe to the nozzle, where it is accelerated and passes into the atmosphere. In a nuclear powered jet engine, the burner section is replaced or supplemented with a nuclear reactor. In the GE design, the air is heated by passing directly over the hot fuel elements. As in any heat exchanger, a certain thermal lag exists between the fuel elements and the air. This results in the instantaneous power generated by the reactor being different than the reactor power delivered to the air. Using conventional jet fuel, the energy is available to the engine as soon as it is burned.

The thermodynamic relationships between the engine components and the reactor can be analyzed in terms of temperatures, airflows and speed. Such parameters as the jet nozzle and compressor stators are assumed to be fixed or stationary.

$$N = f[Q_g, W_a, T_3]\,|_{A_8,\ \beta\ \text{constant}}. \tag{1}$$

In terms of derivatives, the expression for the change in speed due to the actual power delivered to the air, considering this as the only variable, is

$$\Delta N = \frac{dN}{dQ_a}\,\Delta Q_a. \tag{2}$$

Since the power delivered to the air is a function of reactor power generated level, reactor inlet airflow and

* Received by the PGAC, October 26, 1960; revised manuscript received, March 23, 1961.
† Aircraft Nuclear Propulsion Dept., General Electric Co., Cincinnati, Ohio.

inlet temperatures, the change in power is given by

$$\Delta\beta = \Delta A_8 = 0$$

$$\Delta Q_a = \frac{\partial Q_a}{\partial Q_g}\Delta Q_g + \frac{\partial Q_a}{\partial W_a}\Delta W_a + \frac{\partial Q_a}{\partial T_3}\Delta T_3 \tag{3}$$

$$\Delta W_a = \frac{\partial W_a}{\partial N}\Delta N + \frac{\partial W_a}{\partial Q_g}\Delta Q_g \tag{4}$$

$$\Delta T_3 = \frac{\partial T_3}{\partial N}\Delta N + \frac{\partial T_3}{\partial Q_g}\Delta Q_g \tag{5}$$

Substituting (3)–(5) into (2),

$$\frac{\Delta N}{\Delta Q_g} = \frac{\left(\frac{dN}{dQ_a}\right)\left[\left(\frac{\partial Q_a}{\partial Q_g}\right) + \left(\frac{\partial Q_a}{\partial W_a}\right)\left(\frac{\partial W_a}{\partial Q_g}\right) + \left(\frac{\partial Q_a}{\partial T_3}\right)\left(\frac{\partial T_3}{\partial Q_g}\right)\right]}{1 - \left(\frac{dN}{dQ_a}\right)\left[\left(\frac{\partial Q_a}{\partial W_a}\right)\left(\frac{\partial W_a}{\partial N}\right) + \left(\frac{\partial Q_a}{\partial T_3}\right)\left(\frac{\partial T_3}{\partial N}\right)\right]}. \tag{6}$$

The various derivatives are a function of the particular engine and reactor under consideration.

Approximate expressions for the reactor heat transfer process can be derived which are suitable for use in hand calculations. The various approximations which are necessary will be indicated as they occur in the derivation.

The heat energy delivered by the reactor to the airstream is given by

$$Q_a = F(c_{pa}, W_a, \delta T) = c_{pa}W_a\delta T, \tag{7}$$

where δT is the temperature change across a given length of fuel element and c_{pa} is the specific heat of the air.

By assuming a constant specific heat, (7) can be linearized about an operating point.

$$Q_a = \frac{\partial F}{\partial \delta T}\Delta\delta T + \frac{\partial F}{\partial W_a}\Delta W_a. \tag{8}$$

By using zero as a subscript on the variables to denote the partial derivatives evaluated at a point, (8) becomes

$$\Delta Q_a = c_{pa}W_{ao}\Delta\delta T + c_{pa}\delta T_o\Delta W_a, \tag{9}$$

where

$$\frac{\partial F}{\partial \delta T} = c_{pa}W_{ao}; \qquad \frac{\partial F}{\partial W_a} = c_{pa}\delta T_o.$$

In order to investigate how the temperature change across the reactor varies, it is necessary to write the thermodynamic equations describing this behavior:

heat generated = heat stored + heat lost

$$Q_g = M_f c_{pf} T_f(s) + hA(T_f - T_a), \tag{10}$$

where M_f, c_{pf}, T_f, and A refer to the mass, specific heat, surface temperature, and surface area, respectively, of the fuel elements; h is the heat transfer coefficient; T_a is the average, or bulk temperature of the air; and s is the Laplace operator.

A second equation is

heat added to air = heat lost by fuel elements

$$c_{pa}W_a\delta T = hA(T_f - T_a). \tag{11}$$

To derive relatively simple expressions, the assumption is made that the reactor can be considered to be a one-stage fuel element. More elaborate heat transfer studies, both digital and analog, which consider several fuel-element stages, have indicated that this approximation is reasonably valid for the purpose at hand.

With this approximation,

temperature change

= outlet temperature − inlet temperature

$$\delta T = T_4 - T_3. \tag{12}$$

The bulk temperature of the air is then

$$T_a = \frac{T_4 + T_3}{2} = \frac{\delta T}{2} + T_3. \tag{13}$$

The heat transfer coefficient h can be approximated as

$$h = KW_a^{0.8} \tag{14}$$

Further details of the derivation are found in the Appendix. The change in temperature rise across the reactor is given by

$$\Delta\delta T = \frac{\frac{1}{c_{pa}W_{ao}}}{1 + T_R s}\Delta Q_g - \frac{\frac{M_f C_{pf}}{c_{pa}W_{ao}}s}{1 + T_R s}\Delta T_3 - \frac{\frac{Q_{go}}{c_{pa}W_{ao}^2}\left(1 + \frac{0.2M_f C_{pf}}{KAW_{ao}^{0.8}}s\right)}{1 + T_R s}\Delta W_a, \tag{15}$$

where

$$T_R = \frac{M_f C_{pf}}{KAW_{ao}^{0.8}} + \frac{M_f C_{pf}}{2C_{pa}W_{ao}};$$

and the power delivered to the air stream is represented by

$$\Delta Q_a = \frac{1}{(1 + T_R s)}\Delta Q_g - \frac{M_f C_{pf} s}{(1 + T_R s)}\Delta T_3 + \frac{K_a s}{(1 + T_R s)}\Delta W_a. \tag{16}$$

where

$$K_a = \frac{Q_{go}}{W_{ao}} \frac{0.8 M_f C_{pf}}{K A W_{ao}^{0.8}} + \frac{M_f C_{pf}}{2 W_{ao} C_{pa}}.$$

The meaning of (15) and (16) is more clearly understood with a graphical interpretation as presented in Fig. 1.

Returning to (6) and using the partial derivative expressions derived for the reactor with the characteristic expressions for a jet engine,

$$\frac{\partial Q_a}{\partial Q_g} = \frac{1}{1 + T_R s} \qquad \frac{\partial W_a}{\partial Q_g} = K_c$$

$$\frac{dN}{dQ_a} = \frac{K_E}{1 + T_E s} \; K_E > 0 \qquad \frac{\partial T_3}{\partial Q_g} = K_d$$

$$\frac{\partial Q_a}{\partial W_a} = \frac{K_a s}{1 + T_R s} \qquad \frac{\partial T_3}{\partial N} = K_T$$

$$\frac{\partial W_a}{\partial N} = K_W, K_W > 0$$

$$\frac{\partial Q_a}{\partial T_3} = \frac{-K_b s}{1 + T_R s}$$

$$K_b = M_f C_{pf}$$

$$\frac{\Delta N}{\Delta Q_g} = \frac{K_E[1 + (K_a K_c - K_b K_d)s]}{T_E T_R s^2 + (T_E + T_R + K_E K_b K_T - K_E K_A K_W)s + 1.} \quad (17)$$

The system will be stable if both roots of the denominator are negative or have negative real parts. For this to occur, the following inequality exists:

$$T_R + T_E + K_b K_T K_E - K_E K_W K_a > 0. \quad (18)$$

This criterion can be used to test the stability of introducing any type of heat exchanger into a jet engine.

III. Stability Testing of the Heat Transfer Reactor Experiment-3 (HTRE-3) Power Plant

The GE Heat Transfer Reactor Experiment-3 power plant was used to test the criterion developed in (18). The HTRE-3 power plant consists of two X-39-5 turbojet engines to provide cooling airflow for the nuclear reactor. Each engine can operate using either energy from the reactor or from a chemical fuel combustion system. The X-39-5 engine is a major modification of the J-47-GE-11 turbojet engine. The engine has been modified by removal of the combustion section and addition of a compressor discharge scroll, which collects compressor discharge air for ducting to the reactor; and turbine scroll, which takes the hot air from the reactor and/or chemical combustor and distributes it to the turbine inlet annulus. Other major changes include reduction of compressor flow capacity to obtain a suitable match between the compressor and turbine and installation of a new engine control system. The characteristics

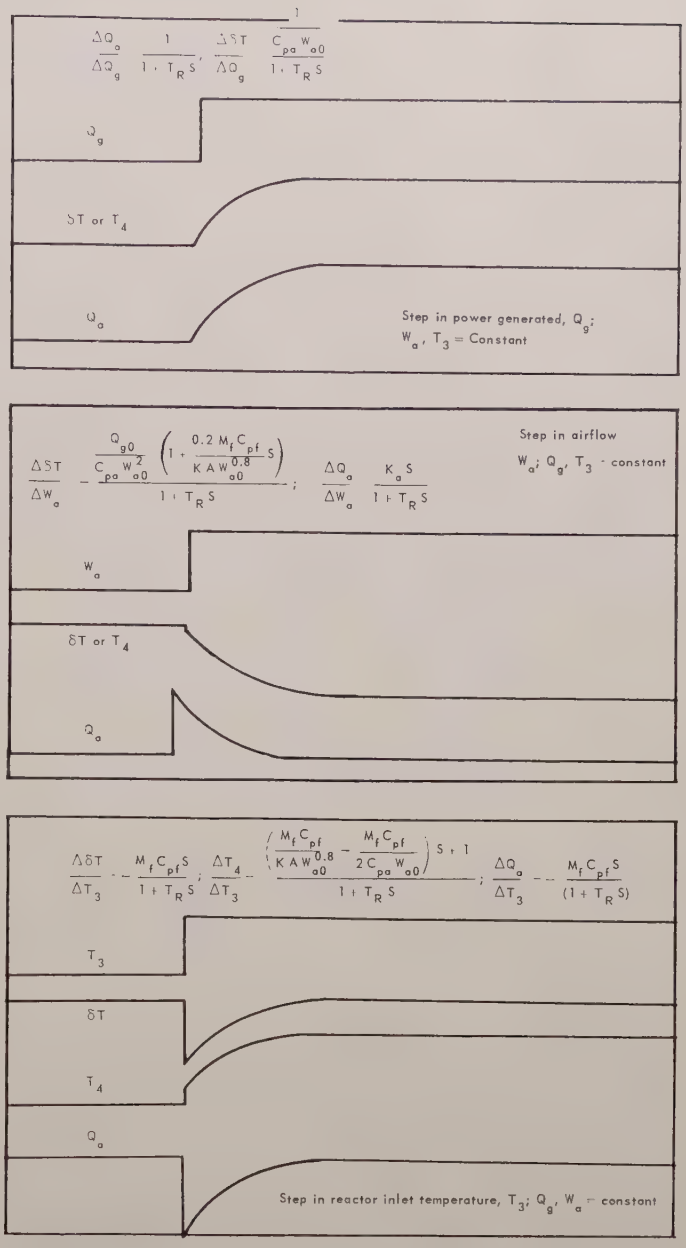


Fig. 1—Effect of varying inlet conditions to the reactor.

curves are typical for a jet engine taking the form shown in Fig. 2.

Data were obtained from the simulated and actual power plants for use in testing the stability criteria developed in (18), where for

$$T_R + T_E + K_b K_T K_E - K_E K_W K_a > 0$$

the power plant will be stable. These data are tabulated in Table I. Various speeds are checked with (18) and the results are given in Table II. It is seen that the stability criteria indicates that the power plant should be stable at 6000 rpm and above, but will be unstable at 5500 rpm and below. Of course, this is still assuming all fixed control parameters, except for varying the reactor power level.

In January, 1960, a test of the HTRE-3 power plant was performed at the GE Idaho Test Facility in order to check the correctness of the analysis. The power

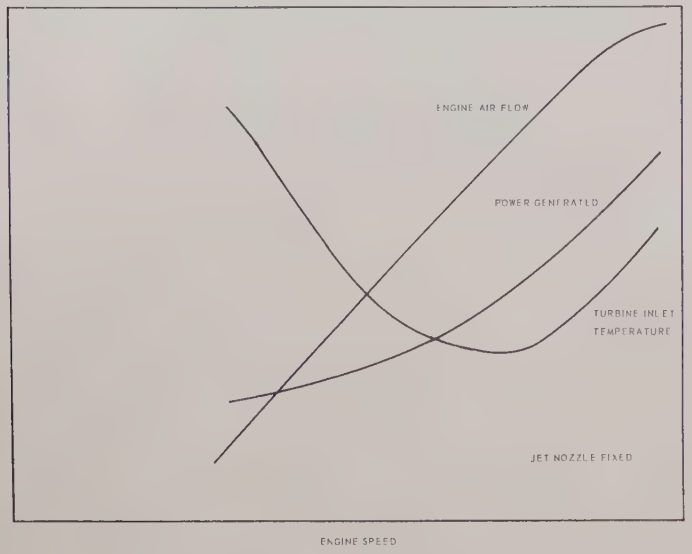


Fig. 2—Steady-state operating characteristics curves for a jet engine.

(Fig. 3). In Fig. 4, 1 per cent power reductions were made starting at 6000 rpm. Above 5900 rpm, the power to air follows the power generated from one operating point to the next, but with a certain time lag. The time lag is the result of the large mass of material in the reactor and the stored thermal energy that must be transferred from the core to the airstream to achieve the new lower turbine inlet temperature. Below 5800 rpm, a decrease in speed requires an increase in temperature for steady-state operation. When power is reduced below 5900 rpm, the temperature starts to decrease, with the resultant decrease in speed and mass flow. The decreasing mass flow tends to raise air temperature to the new operating point requirements, but this requires an increase in reactor core temperatures. To increase the core temperature requires considerable energy and this amount of energy is therefore not available to the airstream. As a result, the power to the airstream decreases, and since the engine cycle operates on Q_a rather

TABLE I
HTRE-3 Power Plant Data

Speed (rpm)	Airflow (No. /sec)	Reactor Power (Mw)	T_R (sec)	T_E (sec)	K_a $\left(\frac{\text{mw}}{\text{No./sec}}\right)$	K_b $\left(\frac{\text{mw}}{^\circ R}\right)$	K_T $\left(\frac{^\circ R}{\text{rpm}}\right)$	K_E $\left(\frac{\text{rpm}}{\text{mw}}\right)$	K_W $\left(\frac{\#/\text{sec}}{\text{rpm}}\right)$
7000	117	23.7	12.5	3	2.52	0.17	0.070	245	0.014
6500	108	21.7	13.4	4	2.70	0.17	0.065	245	0.019
6000	97	19.5	15.0	7	3.01	0.17	0.065	245	0.026
5500	82	17.7	17.7	11	3.84	0.17	0.060	345	0.032
5000	65	16.5	22.3	14	5.66	0.17	0.060	500	0.032

TABLE II
Stability Functions

$$SF = T_R + T_E + K_b K_T K_E - K_E K_w K_a$$

Speed	SF	Remarks
7000	+ 9.8	Stable operating point
6500	+ 7.5	Stable operating point
6000	+ 5.5	Stable operating point
5500	−10.2	Unstable operating point
5000	−49.2	Unstable operating point

plant was successfully operated at a speed of 6000 rpm on a fixed reactor power. When power was stepped down to approximately the value required for 5500 rpm, speed dropped off rapidly. The reactor then had to be shut down so that no damage would occur to the power plant.

To understand further the stability of the cycle, the HTRE-3 was simulated on the analog computer. This simulation indicated that under the conditions imposed, the power plant is stable above 5800 to 5900 rpm. Traces of this are reproduced in Figs. 3 and 4.

Examining these figures closely, the relationship of power generated by the reactor and the power actually delivered to the airstream is seen. The power to the air is computed by $Q_a = W_a C_a \Delta T$. The simulator was initially operated at 100 per cent speed, with a fixed jet nozzle and successive power reductions of 5 per cent than power generated, the further decrease in speed produces an even greater temperature deficiency. Thus, the cycle is unstable below 5800 rpm and confirms the analytical stability analysis.

Another way of understanding the stability of the power plant with the reactor is to look at the speed-energy balance relationships that exist. The characteristic for HTRE-3 is plotted in Fig. 5. The solid line represents the energy required to maintain steady operation at a given speed. If an energy-speed point is above the line, the power plant will accelerate with deceleration occurring below the line.

The energy map is subdivided into four areas using the speed line through the reversal point and the power curves as the boundary lines. Each of these areas will be examined separately for stability.

First, examining Area I, if it is a chemical interburner system, the power delivered to the air always is exactly proportional to fuel flow. Now assume a small perturbation from the steady-state operation line that would increase speed into Area I. The higher speed would result in increased airflow, since

$$\frac{dN}{dt} = K \frac{dW_a}{dt},$$

where K is always positive. Fuel flow is fixed and thus the energy per unit flow drops. The energy is not suffi-

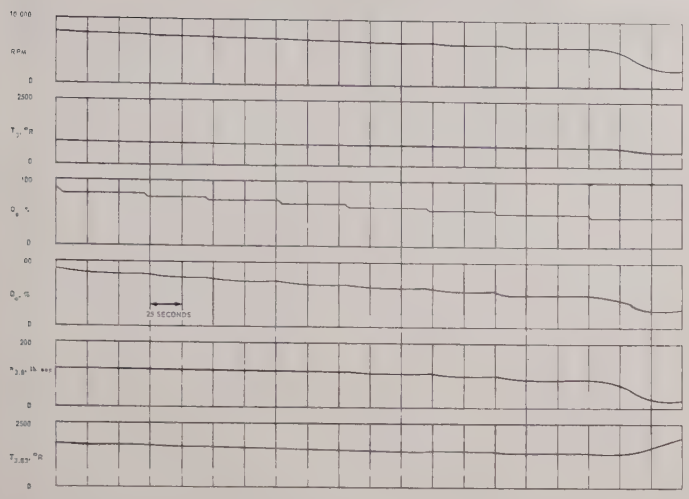


Fig. 3—HTRE-3 power reductions in 5 per cent steps from 7950 rpm, 5000 foot, Standard Day operating conditions, jet nozzle area fixed. The difference between power generated and power delivered to the air stream is shown. Also indicated is the relationship between speed and power generated. The rapid drop-off in speed is seen below 6000 rpm.

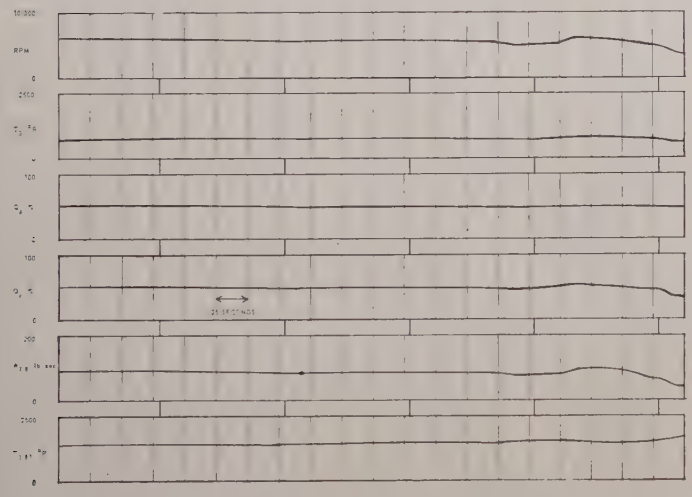


Fig. 4—HTRE-3 power reductions in 1 per cent steps from 6000 rpm, 5000 foot, Standard Day operating conditions, jet nozzle area fixed. Transition from the stable operating region is shown.

cient to maintain the supposed increase in speed and the cycle returns to the original steady-state operating point.

Inserting the reactor in place of the chemical burner, the power generated is the same as the power delivered to the air cycle only during steady-state operation, but unlike the chemical system where fuel flow always is assumed proportional to Q_a, a large thermal lag exists between Q_a and Q_g during transients. With the assumption of a disturbance that increases speed, air flow will increase. Since

$$\frac{\Delta Q_a}{\Delta W_a} = K_A e^{-t/T_R},$$

part of the stored energy from the reactor will initially go into the engine cycle. Thus, Q_a/W_a does not decrease as much as a chemical fuel system where Q_a is fixed, but it will still decrease. Therefore, the energy per unit flow is not available to maintain the hypothesized in-

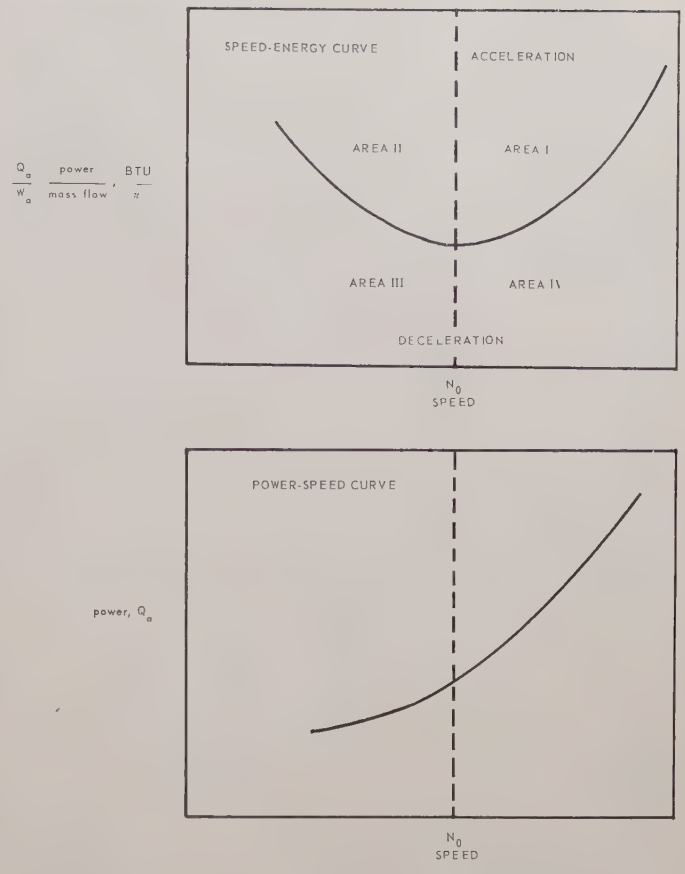


Fig. 5—Power vs speed curves for HTRE-3.

crease in speed and the cycle is therefore stable on nuclear power in Area I.

Turning to Area II, hypothesize once again an increase in speed during chemical operation. As speed increases, the energy per unit flow decreases and with a fixed fuel flow, the increase in airflow will cause W_f/W_a to decrease. However, looking at the power-speed curve it is seen that an increase of fuel flow is required to maintain the new speed, but under the assumption of fixed fuel flow, increased power is not available. Therefore, speed must return to the original operating point.

When the nuclear system is examined, the situation is different. The assumed perturbation increasing speed will require a lower energy per unit flow. Recall that the energy delivered to the engine cycle is now a variable. The greater airflow causes the energy delivered to the cycle by the reactor to increase, but the quantity Q_a/W_a decreases. The additional power required to maintain the higher speed is temporarily available. Eventually, the borrowed stored energy will not be sufficient to maintain the increased speed and, therefore, speed will drop. When the original starting point is approached, however, there is a deficiency in energy available to maintain the operating point and speed drops still further. Thus, the cycle is unstable in Area II.

Examination of Area III indicates the same conclusions would be reached as in Area II with a stable chemical interburner, but unstable on the nuclear reactor. Area IV proves to be stable on both chemical and nuclear power.

IV. Investigation of Control Methods

In previous sections, it has been shown that a nuclear power plant such as HTRE-3 is stable in the upper operating speed range but unstable below this. There are two basic methods available to stabilize the system and extend the operating range of the power plant. The first method is to eliminate the positive feedback loop. The second is to build an automatic feedback control loop around the positive feedback loop such that the over-all system, or exterior loop, is stable.

A. Elimination of the Positive Feedback Loop

The stability criteria developed in Section II stated that for the power plant to be stable the following inequality exists:

$$T_R + T_E + K_b K_T K_E - K_E K_W K_a > 0.$$

Thus, since K_E, K_W, K_a are the gain terms associated with the derivatives of

$$\frac{\partial N}{\partial Q_a}, \quad \frac{\partial W_a}{\partial N}, \quad \frac{\partial Q_a}{\partial W_a},$$

if any of these terms are zero, the system will be stable. Also, the loop can be made stable if any one or any combination of the three can be sufficiently reduced in magnitude. Therefore, a more detailed examination of the partial derivatives is in order.

1) *Discussion of* $\partial N/\partial Q_a$*:* In order to prevent speed from changing with a small change in power to the air, it is necessary to control some parameter in the power plant in such a manner that a desired speed can be maintained. This might be done, for instance, by a variable jet nozzle. Fig. 6 is a typical control parameter map.

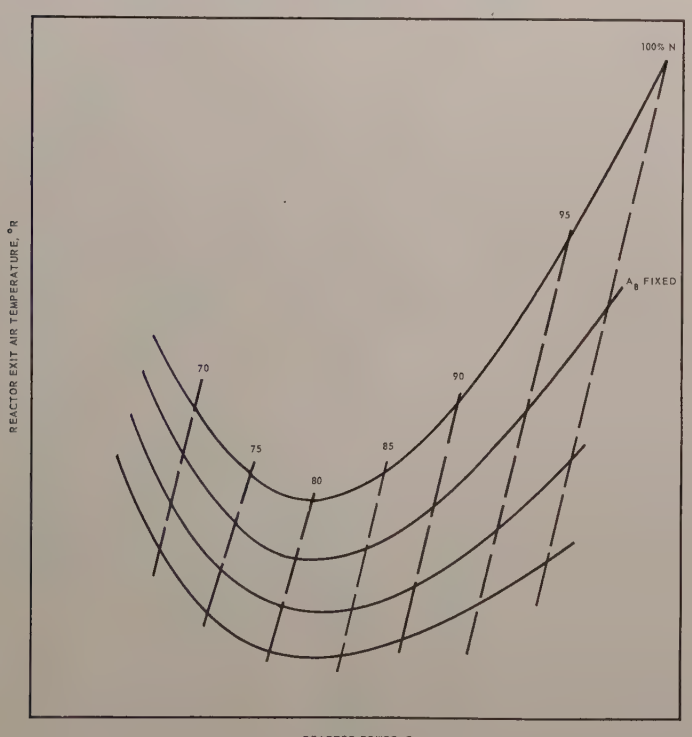


Fig. 6—Control parameter map.

The solid curves are plots of reactor exit air temperature vs reactor power for constant nozzle area, where the reactor power refers to the energy into the air stream. The dashed line curves are plots of reactor exit temperature vs reactor power for constant speed conditions. Since the stability criteria is based on a fixed nozzle, which is represented by the solid lines, it is of interest to examine the dashed line. The slope of these lines is always positive and thus, if the nozzle can be controlled to maintain a fixed speed, $\partial N/\partial Q_a$ can be reduced to zero and the power plant will be stable.

Other methods besides a variable jet nozzle can be used to fix speed around the operating point, but the important thing is that if a parameter is found to control speed, the power plant will be stable.

2) *Discussion of* $\partial W_a/\partial N$*:* The partial derivative $\partial W_a/\partial N$ is determined by the variable geometry of the compressor. The use of an automatic control of the compressor stator blades results in relatively large changes in airflow with changes in speed over the operating range. Fig. 7 shows the relationship of airflow to speed at approximately the scheduled operating point. On the other hand, if the stator levers are controlled by the operator, that is, fixed at each operating point according to some predetermined schedule, the change in airflow with speed about the operating point is relatively small.

An analog computer simulation of a jet engine was arranged so that the compressor airflow was completely independent of speed around an operating point. Admittedly, this was a most optimistic assumption, but it was used as a starting point. The engine cycle was unstable as speed was lowered as a result of power cycle efficiency. At high operating speed, if the airflow is held constant, the amount of power supplied by the reactor must be increased to reach a higher speed. However, around a lower operating speed, less power is needed to obtain a higher speed.

A reactor exit temperature control is not a satisfactory method of controlling the engine cycle in the lower

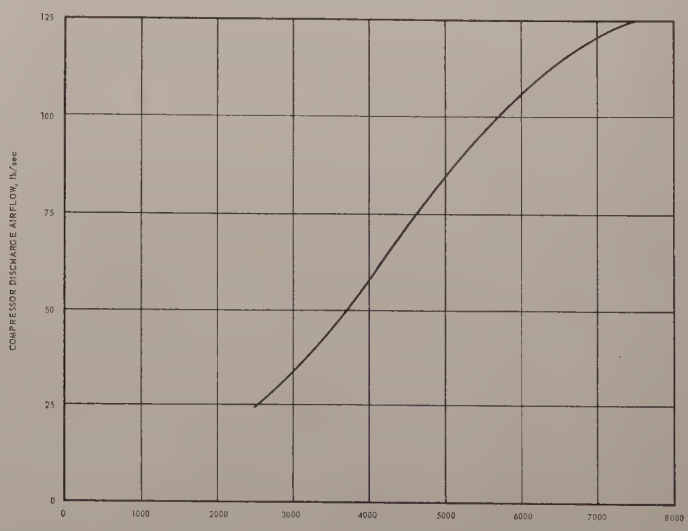


Fig. 7—HTRE-3 approximate airflow vs speed curve.

operating range even with airflow independent of speed because of poorer cycle efficiencies. Difficulties were encountered in the use of an automatic power level control because of the tendency for flux to be a multivalued function of speed, if not restricted to a narrow range. The engine would tend to go to the higher operating point. Therefore, it was concluded that reducing $\partial W_a/\partial N$ was not a feasible solution toward stabilizing the power plant.

3) *Discussion of* $\partial Q_a/\partial W_a$*:* The third partial derivative, $\partial Q_a/\partial W_a$, is determined by the heat transfer characteristics of the reactor. This is a function of the reactor, and in particular a function of the core design. Reducing the time constant of the reactor would reduce this term. This fact was verified with the analog power-plant simulation; reducing the time constant sufficiently permitted operation as desired. However, since the prospect of faster heat transfer from the reactor is not immediately foreseeable, this did not offer a practical method of stabilizing the power plant.

B. Addition of External Engine-Reactor Feedback Paths

It is noted in previous sections that changes in speed also change the instantaneous heat energy delivered to the turbines. Referring to Fig. 8, it is seen that if either ΔQ_{a2} or ΔQ_{a3} changes, ΔQ_{a1} must be made to change in the opposite directions. This requires a change in ΔQ_g, and since $\partial Q_a/\partial Q_g$ represents a lag, the signal to cause the change in ΔQ_g must be an anticipating signal. One such signal is the acceleration of the turbine rotor, obtained either from an accelerometer on the rotor or by differentiation of the engine speed. In theory, then, it should be possible to stabilize the system by putting the rate of change of speed into the flux loop as an additional negative feedback signal.

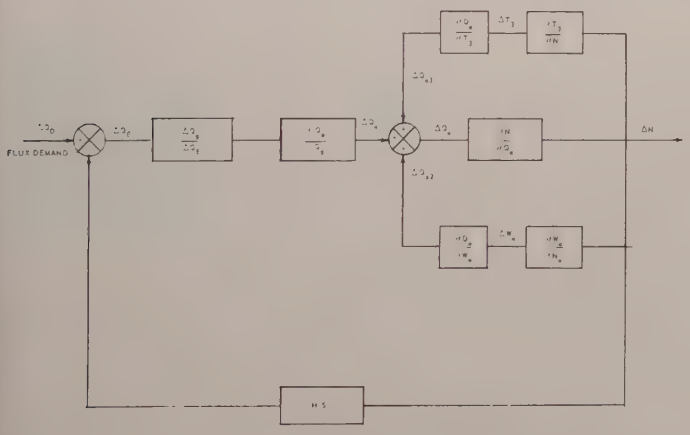


Fig. 8—Nuclear system transfer function block diagram.

The practicability of the solution is another matter. Not indicated in the above discussion is the effect that the rate feedback will have as far as noise is concerned. Also, the system is linearized for operation around a point. Differentiation of signals is generally avoided in control systems as much as possible because of the noise problem. Another aspect is the physical feasibility of obtaining the rate signal.

The rate feedback was introduced into the analog power plant simulator by differentiating the speed signal. A lag was also incorporated into the feedback circuit since a lag would be encountered in the physical system. It was found that the speed range could be increased by this means of stabilization. However, the noise associated with the differentiator was found to keep the reactor control rods in continuous motion. In fact, the noise was such that the rods failed to reset themselves properly. It is therefore doubtful that this means of stabilization would be useful in a practical system.

The development of a speed loop in which engine speed is controlled by reactor power was also considered. This could be accomplished by introduction of a lead network as a forward element of the speed loop in front of the flux loop. This method of stabilizing the power plant offered little encouragement.

The thermodynamic cycle of the engine was studied for other possible methods of stabilization. One possibility from the cycle is the use of the pressure drop across the turbine P_4/P_5, the pressure drop being proportional to $[1+K(dN/dt)]$. Unfortunately, the lead action is not sufficient to stabilize the system and was thus dropped from further consideration.

The introduction of an external engine-reactor feedback path can be used to stabilize the power plant in the form of a speed derivative. However, the noise problem associated with the system would be difficult to solve with physical hardware. Various methods for stabilizing the power plant are summarized in Table III.

TABLE III

SUMMARY OF POWER PLANT STABILITY INVESTIGATION

Stators	Nozzle	Reactor	Remarks
Variable	Fixed	Flux or temperature control	Stable operation above the inflection point in the $T_{3.9c}$ vs Q_a curve. Unstable below the inflection point due to engine-reactor characteristics.
Fixed (assume W_a independent of N)	Fixed	Flux or temperature control	Unstable at lower speeds because of compressor characteristics.
Variable	Automatic	Flux or temperature control	Stable operation over entire range as long as speed loop is dominant.
Variable	Fixed	Flux control with rate feedback of speed	Stable over operating range, but very noisy reactor operation.
Fixed	Automatic	Flux control	Stable operation around fixed schedule.
Variable	Fixed	Flux control with speed over-ride	Unstable at lower end of operating range.
Variable	Fixed	Flux control with feedback of P_4/P_5.	Unstable from initial evaluation because of insufficient lead action.

V. Conclusions

A mathematical relationship has been developed for power plants consisting of a jet engine with a heat exchanger for the power source that can be used to test the stability of such power plants knowing certain partial derivatives. The criterion is based on the assumption that the power level of the heat exchanger can be regulated with all other control parameters being fixed. A test on the HTRE-3 power plant developed by the GE Aircraft Nuclear Propulsion Department indicates the correctness of the analysis.

No difficulty is encountered in operation in the restricted upper-speed region. However, in order to achieve a low idle thrust, it is desirable to extend the speed operating range. Once the effects of inserting a nuclear reactor into the jet engine are understood it is possible to develop power-plant controls to achieve any desired type of operation.

Methods of extending the operating region were investigated, and it was determined that the use of a high-response speed-control system offered the best possibility for achieving the extended operating region. With the speed control it is possible to achieve any desired operation with the nuclear power plant.

Appendix

Eqs. (15) and (16) are derived by linearizing the re-arranged equations of Section II and evaluating the partial derivatives at a given point.

Substituting (11) and (13) into (10), (19) is obtained:

$$Q_g = \frac{M_f c_{pf} c_{pa} W_a \delta T}{hA} s + \frac{M_f c_{pf} \delta T}{2} s + M_f c_{pf} T_3 s + c_{pa} W_a \delta T. \tag{19}$$

Combining (14) and (19), and re-arranging, the following is obtained:

$$\delta T = \frac{Q_g - M_f c_{pf} T_3 s}{\frac{M_f c_{pf} c_{pa} W_a^{0.2}}{KA} s + \frac{M_f c_{pf}}{2} s + c_{pa} W_a} = G(Q_g, T_3, W_a). \tag{20}$$

Eq. (20) may be linearized about an operating point to obtain $\Delta\delta T$:

$$\Delta\delta T = \frac{\partial G}{\partial Q_g} \Delta Q_g + \frac{\partial G}{\partial T_3} \Delta T_3 + \frac{\partial G}{\partial W_a} \Delta W_a. \tag{21}$$

The partial derivatives are obtained from (19). The subscript zero is again used to denote that the partial derivatives are evaluated at a point. After simplification, the results of the differentiation are:

$$\frac{\partial G}{\partial Q_g} = \frac{\frac{1}{W_{ao} c_{pa}}}{\left(\frac{M_f c_{pf}}{KA W_{ao}^{0.8}} + \frac{M_f c_{pf}}{2 c_{pa} W_{ao}}\right) s + 1}$$

$$\frac{\partial G}{\partial T_3} = \frac{-\frac{M_f c_{pf}}{c_{pa} W_{ao}} s}{\left(\frac{M_f c_{pf}}{KA W_{ao}^{0.8}} + \frac{M_f c_{pf}}{2 c_{pa} W_{ao}}\right) s + 1}$$

$$\frac{\partial G}{\partial W_a} = \frac{-\frac{Q_{go}}{c_{pa} W_{ao}^2}\left(1 + \frac{0.2 M_f c_{pf}}{KA W_{ao}^{0.8}} s\right)}{\left(\frac{M_f c_{pf}}{KA W_{ao}^{0.8}} + \frac{M_f c_{pf}}{2 c_{pa} W_{ao}}\right) s + 1}.$$

To simplify the notation, we set

$$\frac{M_f c_{pf}}{KA W_{ao}^{0.8}} + \frac{M_f c_{pf}}{2 c_{pa} W_{ao}} = T_R.$$

Eq. (21), with the evaluated partial derivatives, becomes

$$\Delta\delta T = \frac{\frac{1}{c_{pa} W_{ao}}}{(1 + T_R s)} \Delta Q_g - \frac{\frac{M_f c_{pf}}{c_{pa} W_{ao}} s}{(1 + T_R s)} \Delta T_3 - \frac{\frac{Q_{go}}{c_{pa} W_{ao}^2}\left(1 + \frac{0.2 M_f c_{pf}}{KA W_{ao}^{0.8}} s\right) \Delta W_a}{(1 + T_R s)}. \tag{15}$$

The value of $\Delta\delta T$ obtained from (15) can now be substituted into (9) with the following results:

$$\Delta Q_a = \frac{1}{(1 + T_R s)} \Delta Q_g - \frac{M_f c_{pf} s}{(1 + T_R s)} \Delta T_3 + \frac{\frac{Q_{go}}{W_{ao}}\left(0.8 \frac{M_f c_{pf}}{KA W_{ao}^{0.8}} + \frac{M_f c_{pf}}{2 W_{ao} c_{pa}}\right) s \Delta W_a}{(1 + T_R s)} \tag{22}$$

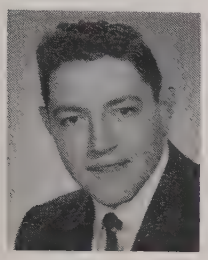
$$\frac{Q_{go}}{W_{ao}}\left(\frac{0.8 M_f c_{pf}}{KA W_{ao}^{0.8}} + \frac{M_f c_{pf}}{2 W_{ao} c_{pa}}\right) = K_a.$$

Then (22) becomes:

$$\Delta Q_a = \frac{1}{(1 + T_R s)} \Delta Q_g - \frac{M_f c_{pf} s}{(1 + T_R s)} \Delta T_3 + \frac{K_a s}{(1 + T_R s)} \Delta W_a. \tag{16}$$

Contributors

Michael Athanassiades (S'58) was born in Drama, Greece on May 3, 1937. He received the B.S.E.E. and M.S.E.E. degrees in 1958 and 1959, respectively, from the University of California at Berkeley. At present, he is a graduate student at the University of California working toward the Ph.D. degree in electrical engineering.



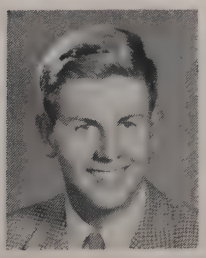
M. Athanassiades

He has taught at the University of California since 1958, and has conducted research in the fields of linear and nonlinear feedback control systems. At present, he is part-time research consultant to the M.I.T. Lincoln Laboratories, Lexington, Mass., in the field of adaptive control systems. He has published research in the field of nonlinear control systems.

Mr. Athanassiades is a member of Phi Beta Kappa, Eta Kappa Nu, Sigma Xi, and a student member of AIEE.

❖

Peter K. Bohacek was born in Olomouc, Czechoslovakia, on March 4, 1936. He received the B.E.E. and M.E.E. degrees from Yale University, New Haven, Conn., in 1958 and 1959, respectively. He is now working toward the Ph.D. degree at Yale in the field of information theory.



P. K. Bohacek

During the summers of 1959 and 1960, he was employed by Sylvania Electric Products, Inc., Amherst, N. Y., where he worked on pulse-shaping analysis and multipath communication.

Mr. Bohacek is a member of Tau Beta Pi.

❖

David Buden (M'59) was born in Philadelphia, Pa., on July 10, 1930. He received the B.S. degree in 1952 and the M.S. degree in geophysics in 1954, from the Pennsylvania State University, State College, Pa., and has done graduate work in mathematics at the University of Cincinnati, Ohio.

D. Buden

Since 1955 he has been with the Aircraft Nuclear Propulsion Department of the General Electric Company, Cincinnati, Ohio, engaged in the transient analysis of nuclear power plants. This includes the integration of the nuclear power plant's chemical, nuclear and auxiliary control systems to assure their compatibility, safe operation and optimum performance. Also included is work on the development of advanced power plant control concepts, and operational philosophies. He is presently the Supervisor of the Control System Analysis Unit, with the responsibility for the transient analysis and analog simulation of power plants and associated control systems.

Mr. Buden is a member of the Midwestern Simulation Council and Sigma Gamma Epsilon.

❖

Jose B. Cruz, Jr. (S'56–M'57) was born in Bacolod City, Philippine Islands, on September 17, 1932. He received the B.S.E.E. degree *summa cum laude* from the University of the Philippines, Quezon City, in 1953, the M.S.E.E. degree from the Massachusetts Institute of Technology, Cambridge, in 1956, and the Ph.D. degree in electrical engineering from the University of Illinois, Urbana, in 1959.

J. B. Cruz, Jr.

After one year of teaching at the University of the Philippines, he went to M.I.T. in 1954 as research assistant in the Research Laboratory of Electronics. In 1956, he joined the faculty of the University of Illinois as instructor in electrical engineering. Since September, 1959, he has been assistant professor of electrical engineering and research assistant professor with the Coordinated Science Laboratory, University of Illinois.

Dr. Cruz is a member of Sigma Xi, Phi Kappa Phi, and AAUP.

❖

Charles A. Desoer (S'50–A'53–SM'57), for a photograph and biography, please see page 93 of the February, 1961, issue of these Transactions.

❖

Herbert L. Groginsky (S'51–A'53–M'58) was born in Newark, N. J., on July 10, 1930. He received the B.E.E. degree from the Polytechnic Institute of Brooklyn, N. Y. in 1952, and the M.S. and Eng. Sc.D. degrees in electrical engineering from Columbia University, New York, N. Y., in 1952 and 1959, respectively. From 1951 to 1952, he served as a teaching assistant in the physics department at department at Polytechnic Institute of Brooklyn.

From 1952 to 1959, he was a member of the staff of the Electronics Research Laboratories of Columbia University, where he engaged in radar systems analyses and studied adaptive control systems. Since June, 1959, he has been associated with the Advanced Development Laboratory of the Raytheon Company, Waltham, Mass., where he is engaged in research on problems in communication and information theory. He also currently teaches a course in radar systems engineering at Northeastern University, Boston, Mass.

H. L. Groginsky

Dr. Groginsky is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

❖

Dale R. Ingwerson (M'56) was born in Lodge Pole, Neb., on March 22, 1925. In 1950 and 1951, respectively, he received the B.S. and M.A. degrees in physics from the University of Nebraska, Lincoln.

D. R. Ingwerson

He began part-time course work in applied mechanics at the Polytechnic Institute of Brooklyn, Brooklyn, N. Y., which he continued at Stanford University, Stanford, Calif., in 1956. In January, 1961, he received the Ph.D. degree from Stanford with a dissertation on nonlinear control theory. From 1951 to 1953, he was employed as an aeronautical engineer by Sverdrup and Parcel Inc., St. Louis, Mo. In 1953, he joined the Sperry Gyroscope Company in Great Neck, N. Y. Since then he has been engaged in the design of servomechanisms, computers, and inertial devices. In 1956, he was transferred to Sunnyvale, Calif.

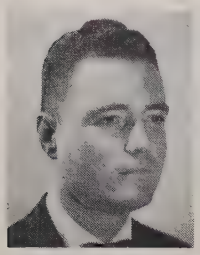
Dr. Ingwerson is a member of Sigma Xi and Pi Mu Epsilon.

❖

Robert B. Kerr (M'59) was born in Pittsburgh, Pa., on February 11, 1929. He received the B.S. degree from Lafayette College, Easton, Pa., in 1950, the M.S. degree from the Massachusetts Institute of Technology, Cambridge, in 1955, and the D.Eng. degree from the Johns Hopkins University, Baltimore, Md., in 1959, all in electrical engineering. In 1958–1959, he was a research staff assistant in the area of signal

detection and analysis for the Johns Hopkins Radiation Laboratory.

He has been an assistant professor of electrical engineering at Princeton University, Princeton, N. J., since 1959. His industrial experience has included work in the Flight Test Division of the Boeing Aircraft Company, Seattle, Wash., and consulting work for the Seismograph Service Corporation, Tulsa, Okla., for the analysis of seismic signals. He is a consultant to the Aeronautical Research Associates of Princeton, Inc., for the study of data analysis problems associated with control systems.

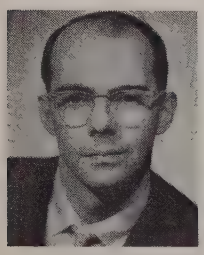


R. B. KERR

Dr. Kerr is a member of Sigma Xi and the AIEE.

❖

W. Wayne Lichtenberger (S'55–M'61) was born in Carmi, Ill., on September 25, 1933. He received the B.S. degree in 1955, the M.S. degree in 1956, and the Ph.D. degree in 1961, in electrical engineering all from the University of Illinois, Urbana.



W. LICHTENBERGER

While engaged in graduate work, he held the positions of research assistant in the Electrical Engineering Research Laboratories, where he worked on cathodic protection, instructor in the Department of Electrical Engineering, and research associate in the Coordinated Science Laboratory, working in high-resolution radar and adaptive control systems. He is now research assistant professor of electrical engineering in the Department of Electrical Engineering and in the Coordinated Science Laboratory, currently engaged in research in the area of teaching machines.

Dr. Lichtenberger is a member of Sigma Xi, Tau Beta Pi, Phi Kappa Phi, Eta Kappa Nu, and Pi Mu Epsilon.

❖

Richard J. McGrath (S'57–M'60), for a photograph and biography, please see page 93 of the February, 1961, issue of these TRANSACTIONS.

❖

Robert F. Miller was born in Salamanca, N. Y., on May 20, 1931. He received the B.S.E.E. degree from Northwestern University, Evanston, Ill., in 1953, and the M.S. degree in mathematics from Xavier University, Cincinnati, Ohio, in 1961.

R. F. MILLER

From 1953 to 1957, he served in the U. S. Navy, the last two years as a member of the faculty of the Marine Engineering Department of the U. S Naval Academy at Annapolis, Md.

Since 1957, he has been employed by the Aircraft Nuclear Propulsion Department of the General Electric Company, Cincinnati, Ohio, working on design and development of power plant control systems, with emphasis on reactor control systems.

Mr. Miller is an associate member of the AIEE.

❖

Gordon J. Murphy (M'55) was born in Milwaukee, Wis., on February 16, 1927. He received the B.S.E.E. degree from the Milwaukee School of Engineering in 1949, the M.S.E.E. degree from the University of Wisconsin, Milwaukee, in 1952, and the Ph.D. degree from the University of Minnesota, Minneapolis, in 1956.

G. J. MURPHY

He held a position as Assistant Professor of electrical engineering at the Milwaukee School of Engineering from 1949 to 1951, at which time he accepted a position as Project Engineer on inertial guidance systems at the AC Spark Plug Division of General Motors in Milwaukee. In 1952, he accepted a position as Instructor in electrical engineering at the University of Minnesota, where he developed and taught courses in automatic control while continuing his studies as a part-time student. In 1956, he was appointed Assistant Professor of electrical engineering at the University of Minnesota.

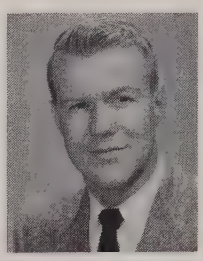
In 1957, he became an Associate Professor of electrical engineering in the Technological Institute of Northwestern University, Evanston, Ill. He has been engaged since then in teaching and in research in the fields of statistical control theory, sampled-data theory, and adaptive control. Since 1960, he has been Professor and Chairman of the Department of Electrical Engineering at Northwestern.

Dr. Murphy is a member of AIEE, ASEE, Sigma Xi, and Eta Kappa Nu.

❖

Eugene A. O'Hern was born in Flint, Mich., on January 20, 1927. He received the B.S.M.E., M.S.M.E. and Ph.D. degrees from Purdue University, Lafayette, Ind., in 1948, 1949 and 1951, respectively.

He joined the staff of North American Aviation Inc., in 1951, as a research engineer in the field of flight control analysis, working on advanced state-of-the-art flight control concepts on such important missile projects as the X-10 and XSM-64(NAVAHO). In 1954, he directed the flight control activity on the NAVAHO program. In 1957, he engaged in preliminary design activity of avionics systems. In 1961, he was appointed Assistant Chief of the Preliminary Engineering Section, Armament and Flight Control Division, with responsibilities for analysis, synthesis, and preliminary design of flight control systems.



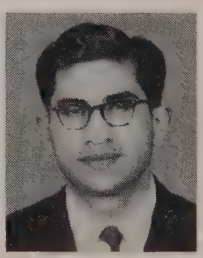
E. A. O'HERN

Dr. O'Hern is a member of the University of Southern California, graduate school faculty, and has organized and taught evening courses in aircraft dynamics at that school.

He is a member of Pi Tau Sigma, Tau Beta Pi, and Sigma Xi.

❖

Narindra N. Puri was born in New Delhi, India, on November 30, 1933. He received the B. Tech. degree in electrical engineering from the Indian Institute of Technology, Kharagpur, in 1955.



N. N. PURI

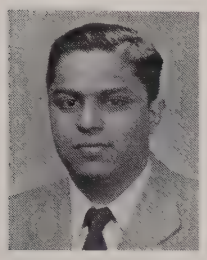
From 1955 to 1957, he worked with Brown Boverie and Cie. in Germany and Switzerland. In 1957, he became a research assistant at the University of Wisconsin, Madison. He was with the Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, as a Harrison Research Fellow, from 1958 until 1960, and received the Ph.D. degree there in 1961.

At present, Dr. Puri is a member of the faculty of Drexel Institute of Technology, Philadelphia.

❖

Vaidyeswaran Rajarman (S'58) was born in Madras, India, on September 8, 1933. He received the B.Sc. (Hons.) degree in physics from the University of Delhi, Delhi, India, in 1952. He received the Diploma and the Associateship in electrical communications engineering from the Indian Institute of Science, Bangalore, in 1955 and 1957, respectively. He was awarded an overseas

scholarship by the Government of India in 1957, and obtained the M.S. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1959. He is currently working towards his doctorate at the University of Wisconsin, Madison.



V. Rajaraman

Mr. Rajaraman is a member of Sigma Xi.

❖

Zenonas V. Rekasius was born in Panevezys, Lithuania, on January 1, 1928. He received the B.S.E.E. and M.S.E.E. degrees from Wayne State University, Detroit, Mich., in 1954 and 1956, respectively, and the Ph.D. degree in electrical engineering from Purdue University, Lafayette, Ind., in 1960.

Z. V. Rekasius

From 1954 to 1956, he held a graduate teaching assistantship at Wayne State University. From 1956 to 1957, he was assistant professor of electrical engineering at the Detroit Institute of Technology; and from 1957 to 1960, he served as instructor of electrical engineering and research assistant in the Statistical and Computer Laboratory of Purdue University. He is presently assistant professor of electrical engineering at Purdue.

Dr. Rekasius is an associate member of AIEE.

❖

Vincent C. Rideout (M'44–SM'53–F'60), for a photograph and biography, please see page 94 of the February, 1961, issue of these Transactions.

❖

Kanji Sahara was born in Hiroshima, Japan, on April 4, 1934. He received the B.S.E.E. degree in 1956 from Illinois Institute of Technology, Chicago, and the M.S. and Ph.D. degrees in 1959 and 1961, respectively, from Northwestern University, Evanston, Ill.

K. Sahara

He worked for Sperry Gyroscope Company, Great Neck, N. Y., from 1956 to 1957.

Dr. Sahara is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

Otto J. M. Smith (M'44–SM'51–F'59) was born in Urbana, Ill. on August 6, 1917. He received the B.S. degree in chemistry from Oklahoma Agricultural and Mechanical College, Stillwater, and the B.S.E.E. degree from the University of Oklahoma, Norman in 1938. He was a research assistant at the H. G. Ryan High Voltage Laboratory at Stanford University, Stanford, Calif., from 1938 until 1941, when he received the Ph.D. degree in electrical engineering.

O. J. M. Smith

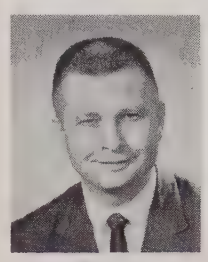
He has taught and conducted research at Tufts College, Medford, Mass., Doble Engineering Company, Medford, Mass., Denver University, Colo., Westinghouse Research Laboratory, Summit Research and Development Laboratory, Scranton, Pa., Shell Development Company, Emeryville, Calif., Radio Interference Specialty Company, San Francisco, Calif., and Donner Scientific Company, Concord, Calif. He has been at the University of California in Berkeley since 1947. From 1954 to 1956, while on leave from the University of California, he was a visiting professor of servomechanisms at the Instituto Tecnologico de Aeronautica, Sao Jose dos Campos, Estado de Sao Paulo, Brazil. He studied the educational system of Brazil, made recommendations, and taught advanced automatic control systems for the Brazilian government. As a Guggenheim Fellow in 1960, he conducted research at the Institut für Regelungstechnic, Technische Hochschule Darmstadt, West Germany, and traveled extensively in Russia. He was a member of the American Automatic Control Council inspection team for the Japanese Government in May, 1961. He is presently a professor of electrical engineering at the University of California in Berkeley.

He is the inventor of a low-frequency sine-function generator, an X-ray thickness gauge, low-frequency test equipment, deadbeat predictor controls for automatic systems, a constant-frequency variable-speed generator, a controlled-torque ac motor, and a phase-shift scaler. He has published research in the fields of magnetic amplifiers, magnetic frequency multipliers, semiconductors, phonograph recording, high-voltage corona, power-line fault locators, radiation instrumentation, counters, education, economic analogs, nonlinear feedback systems, machinery and statistics.

Dr. Smith is a Fellow of the AAAS, and a member of the AIEE, the ASEE, American Automatic Control Council, American Physical Society, American Institute of Physics, Sigma Xi, Phi Kappa Phi, Tau Beta Pi, Eta Kappa Nu, Phi Lambda Upsilon, Alpha Phi Omega, Kappa Tau Pi, and Phi Eta Sigma.

❖

Richard K. Smyth (A'52–M'57) was born in Ada, Okla., on August 30, 1929. He received the B.S. degree in physics from the California Institute of Technology, Pasadena, in 1951. In 1958, he received the M.S.E.E. degree from the University of Southern California, Los Angeles, where he is presently a candidate for a Ph.D. degree in electrical engineering, specializing in servo theory, computers, circuits and mathematics.



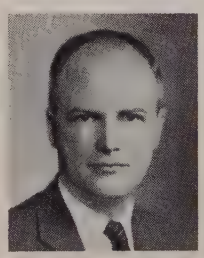
R. K. Smyth

From 1951 to 1955, he worked at Statham Laboratories, Los Angeles, and for Wiancko Engineering Company, Pasadena, in the field of dynamic instrumentation and electronics. He joined Autonetics, a division of North American Aviation, Inc., Downey, Calif., in 1955 and has successively held the positions of research engineer, senior research engineer, and engineering supervisor in the Armament and Flight Control Engineering Division. He has worked in the area of synthesis and analysis of advanced flight, control systems and landing systems for vehicles such as the F-107, F-108, B-70, A3J, GAM-77A, NAVAHO, F-102, and DYNA SOAR.

Mr. Smyth is a lecturer in circuit synthesis and electromechanics in the electrical engineering department at the University of Southern California.

❖

William H. Surber, Jr., (M'47) was born in Charlottesville, Va., on April 14, 1920. He received the B.S. degree from the University of Richmond, Richmond, Va., in 1941. He received the E.E. degree in 1943, and the Ph.D. degree in electrical engineering in 1948, from Princeton University Princeton, N. J.



W. H. Surber

From 1944 to 1946, he served in the U. S. Navy as a member of the Electronic Field Service Group at the Naval Research Laboratory in Washington, D. C. In 1948, he was appointed an assistant professor at Princeton University, and he became an associate professor in 1951. He has been a professor of electrical engineering since 1956.

He has been a consultant to the Brookhaven National Laboratory, Upton, N. Y., to the Electronics Division of the Curtiss-Wright Corporation, Carlstadt, N. Y., for the development of aircraft simulators and analog computing systems, and to a number of other companies in the fields of electronic circuits and feedback control systems. At present he is a senior consultant and director of the Aeronautical Research Associates of Princeton, Inc., for work in the area of adaptive control systems.

Dr. Surber is a member of Phi Beta Kappa Sigma Xi, the AIEE, and the American Physical Society.

Albert I. Talkin (A'50–M'55) was born in Brooklyn, N. Y., on January 3, 1924. He received the B.A. degree in physics from Brooklyn College in 1944, and the M.A. degree in mathematics from Columbia University, New York, N. Y., in 1947.

A. I. Talkin

From 1944–1946, he served as an army electronics technician on the Manhattan Project, Los Alamos, N. M. In 1947, he joined the Rotating Physics Program at General Electric Company, Schenectady, N. Y., where, from 1948–1950, he did circuit design work in the receiving tube department. From 1950 to 1953, he was with the National Bureau of Standards, Washington, D. C. He became a member of the Diamond Ordnance Fuze Laboratories, Washington, D. C., at its activation in 1953. His major experience has been in circuit design and feedback control, and at present he is Research Supervisor in the Countermeasures and Special Systems Branch.

❖

Franz B. Tuteur (S'49–A'51–M'56) was born in Frankfurt-am-Main, Germany, on March 6, 1923. He received the B.S. degree from the University of Colorado, Boulder, in 1944, and the M.E. and Ph.D. degrees from Yale University, New Haven, Conn., in 1949 and 1954, respectively, all in electrical engineering. Since 1950 he has been a full-time staff member at the Electrical Engineering Department of Yale and currently holds the rank of Associate Professor. His major fields of interest are in servomechanisms and in communications theory, and he has been active in a number of research projects in these fields. He is co-author of the textbook "Control System Components."

F. B. Tuteur

Dr. Tuteur is a member of Tau Beta Pi, Sigma Xi, and Eta Kappa Nu.

❖

Gerald E. Tutt was born in Richmond, Va., on April 6, 1935. He received the B.S. degree in mechanical engineering from the University of California, Berkeley, in 1958, and has completed work towards the M.S. degree in mechanical engineering at the University of Southern California, Los Angeles.

G. E. Tutt

In 1958, he joined the Douglas Aircraft Company, Santa Monica, Calif., where he worked on the engine servo analysis of the Able II second stage. Subsequently, his duties involved the control system analysis of the Able III and IV first stage, and the Transit-Courier first stage. From June–September, 1960, he served as the Douglas Guidance and Control Section representative at Vandenburg Air Force Base, Santa Maria, Calif. In this capacity, he assisted with the autopilot checkout and the preliminary flight analysis of the Discoverer space vehicle. At present, he has been investigating control techniques for flexible vehicles.

Mr. Tutt is a member of the ASME.

❖

Walter K. Waymeyer (M'60) was born in Covington, Ky., on August 20, 1927. He received the B.S.M.E. degree from the University of Cincinnati, Cincinnati Ohio, and the M.S.M.E. degree from the University of Kansas, Lawrence, in 1949 and 1951, respectively. Since then he has studied electronics and automatic control at Cornell University, Ithaca, N. Y., New Mexico State University, University Park, and the University of California at Los Angeles.

In 1951, as a lieutenant in the U. S. Army Ordnance, he was assigned to the White Sands Missile Range, White Sands, N. M., where he became associated with the development and testing of guided missile control systems. In 1954, he joined the Douglas Aircraft Company, Santa Monica, Calif., where he has had control design responsibility in the Nike, Thor, Skybolt and Saturn programs. Since 1959, he has held his present position as Supervisor of the Control Group.

W. K. Waymeyer

Mr. Waymeyer is a member of Pi Tau Sigma and Tau Beta Pi.

❖

Cornelius N. Weygandt was born in Germantown, Pa., on August 13, 1904. He received the B.S.E.E. degree from the Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, in 1928, the M.S.E.E. degree from the Massachusetts Institute of Technology, Cambridge, in 1933, and the Ph.D. from the Moore School, in 1937.

C. N. Weygandt

From 1928 to 1932, he was employed in various departments of the General Electric Company, and for two years thereafter, he was with the Brooke Engineering Company, Philadelphia. He then joined the staff of the Moore School, where he is presently Professor of Electrical Engineering, teaching courses in Energy Conversion and in Control Instrumentation.

❖

Jack Wing (S'51–A'52–M'57), for a photograph and biography, please see page 94 of the February, 1961, issue of these TRANSACTIONS.